RESEARCH LETTER

Open Access

The effectiveness of machine learning methods in the nonlinear coupled data assimilation

Zi-ying Xuan[1,3], Fei Zheng[1,2*] and Jiang Zhu[1,3]

Abstract

Implementing the strongly coupled data assimilation (SCDA) in coupled earth system models remains big challenging, primarily due to accurately estimating the coupled cross background-error covariance. In this work, through simplified two-variable one-dimensional assimilation experiments focusing on the air–sea interactions over the tropical pacific, we aim to clarify that SCDA based on the variance–covariance correlation, such as the ensemble-based SCDA, is limited in handling the inherent nonlinear relations between cross-sphere variables and provides a background matrix containing linear information only. These limitations also lead to the analysis distributions deviating from the truth and miscalculating the strength of rare extreme events. However, free from linear or Gaussian assumptions, the application of the data-driven machine learning (ML) method, such as multilayer perceptron, on SCDA circumvents the expensive matrix operations by avoiding the explicit calculation of background matrix. This strategy presents comprehensively superior performance than the conventional ensemble-based assimilation strategy, particularly in representing the strongly nonlinear relationships between cross-sphere variables and reproducing long-tailed distributions, which help capture the occurrence of small probability events. It is also demonstrated to be cost-effective and has great potential to generate a more accurate initial condition for coupled models, especially in facilitating prediction tasks of the extreme events.

Keywords Coupled data assimilation, Coupled cross error covariance, Nonlinear relationship, Machine learning, Extreme events

Introduction

As more record-breaking weather events occur under global warming, using coupled earth system models to produce reliable seasonal to decadal predictions is progressively crucial for decision-makers to manage the risk (Wang et al. 2017b; Penny and Hamill 2017; Raymond et al. 2020). The accuracy of model initialization is significant for predictions on seasonal to decadal timescales (Boer et al. 2016). Coupled data assimilation (CDA) serves as a solution, which combines the prior model predictions and observations from different earth components together to obtain the best initial conditions for each component. By maintaining the interaction between different components, CDA mitigates the initial shock and generates physically balanced initial conditions (Zhang 2011; He et al. 2020). Different from weakly CDA (WCDA), strongly CDA (SCDA) allows observations directly influence the state estimation of another component through coupled cross background-error covariance (CCEC). The distinguished performance of SCDA has

*Correspondence:
Fei Zheng
zhengfei@mail.iap.ac.cn
[1] International Center for Climate and Environment Science (ICCES), Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029, China
[2] Collaborative Innovation Center On Forecast and Evaluation of Meteorological Disasters, Nanjing University of Information Science and Technology, Nanjing 210044, China
[3] University of Chinese Academy of Sciences, Beijing 100049, China

been proved by models of various complexities (Zheng and Zhu 2010; Park et al. 2015; Sluka et al. 2016; Penny et al. 2019; Kalnay et al. 2023). While SCDA is a theoretically optimal approach, current operational centers mostly combine the existing atmospheric and oceanic assimilation systems to construct WCDA systems (Fujii et al. 2021). This is because the incorrect CCEC in SCDA will lead to inferior analysis quality compared to WCDA (Han et al. 2013).

It remains challenging for estimating accurate CCEC and implementing SCDA, due to the difference in variability of each component, lead–lag correlation between components, sampling errors, high computational cost, and nonlinear interaction at the interface (Penny et al. 2017; Zhang et al. 2020; Zheng et al. 2022). Attempts have been made to surmount these difficulties, and one such effort involves the development of a multi-timescale, high-efficiency approximate EnKF (MSHea–EnKF). It aims to enhance the computational efficiency and the accuracy of error statistics for slow scale (Yu et al. 2019), while the high observation frequency of fast scale can also help address the multiscale problem (Tondeur et al. 2020). Leading averaged coupled covariance (LACC) method is proposed to alleviate issues arising from lead–lag relationships, and the real-world assimilation experiments also prove that LACC could produce high-quality analyses (Lu et al. 2015; Sun et al. 2020). Reconditioning, Schur product localization (Smith et al. 2018) and the correlation–cutoff method (Yoshida and Kalnay 2018) are three effective approaches for mitigating the sampling error. Although proven to be powerful in state estimation, the dominant SCDA methods, including ensemble-based, variational and hybrid frameworks, have disadvantages of expensive computational cost and assumptions on linearity and Gaussianity that are detrimental to complex high-dimensional coupled models (Zhang and Zhang 2012; He et al. 2017; Evensen et al. 2022). Even the particle filter (PF), free from linear or Gaussian assumptions, faces the unavoidable curse of dimensionality and filter degeneracy when applied to geophysical systems with high dimensions. Some PF variants have been proposed to mitigate these problems by introducing localization schemes or giving equal particle weights (Tödter and Ahrens 2015; Poterjoy 2016; Zhu et al. 2016; Skauvold et al. 2019; Feng et al. 2020).

The data-driven machine learning (ML) method has drawn tremendous attention today, due to its capability of nonlinear expression and spatiotemporal feature extraction, coupled with the advantages of strong generalization and computational efficiency (Sarker 2021; Xu et al. 2021). The successful applications of ML in assimilation for single models provide a promising approach for addressing the aforementioned challenges (Brajard et al. 2020; Arcucci et al. 2021; Ruckstuhl et al. 2021; Huang et al. 2021; Zhou and Zhang 2023). This study aims to investigate (1) the limitation of conventional SCDA strategy that based on variance–covariance correlation and (2) the potential effectiveness of ML in nonlinear SCDA. This paper is organized as follows: Sect. "Problem Definition" elucidates the conflict between the linear update mechanism of conventional assimilation methods and the nonlinear reality. The nonlinear assimilation experiment design is presented in Sect. "Experiment settings for strategy effectiveness evaluation" and the main results are analyzed in Sect. "Performance of different SCDA strategies". Conclusions will be provided in Sect. "Conclusion and discussion".

Problem definition

Data sets

The present study employs the monthly reanalysis data of oceanic and atmospheric components: the sea surface temperature (SST) is obtained from the Hadley Centre Sea Ice and SST data set version 1 (HadISST1) (Rayner et al. 2003); the sea surface salinity (SSS) is from the Hadley Centre's subsurface temperature and salinity data set EN4.2.2 (Gouretski and Reseghetti 2010); and the sea surface height (SSH) measurements is derived from the Simple Ocean Data Assimilation product (SODA 3.15.2) (Carton et al. 2018). The outgoing longwave radiation (OLR) is from NOAA interpolated outgoing longwave radiation data set (Liebmann and Smith 1996); the precipitation rate (PRC) is taken from the Global Precipitation Climatology Project (GPCP) (Adler et al. 2003); and the air temperature at 2 m (T2m) is from the fifth generation ECMWF Reanalysis (ERA5) (Hersbach et al. 2020). The SSH data is available from 1980 to 2020, while others cover the period from 1979 to 2022. The provided data will be used to evaluate the performance of different SCDA strategies in handling relations of various complexity.

Linear characteristic of coupled data assimilation

Ensemble Kalman filter (EnKF) has shown powerful capability in SCDA for coupled ocean–atmosphere models (Liu et al. 2013). The analysis equation of EnKF can be written as (Sakov and Sandery 2015):

$$\boldsymbol{X}^a = \boldsymbol{X}^p + K\left(\boldsymbol{X}^o - \boldsymbol{H}\boldsymbol{X}^p\right) \tag{1}$$

where superscripts a, p and o represent analysis, prediction and observation, respectively, hereafter; $\boldsymbol{X}^a$ represents the analysis field; $\boldsymbol{X}^p$ denotes the background field (a.k.a. the prior model prediction field); $\boldsymbol{X}^o$ is the observation field; $\boldsymbol{H}$ is the linearized observation operator; $\boldsymbol{K}$

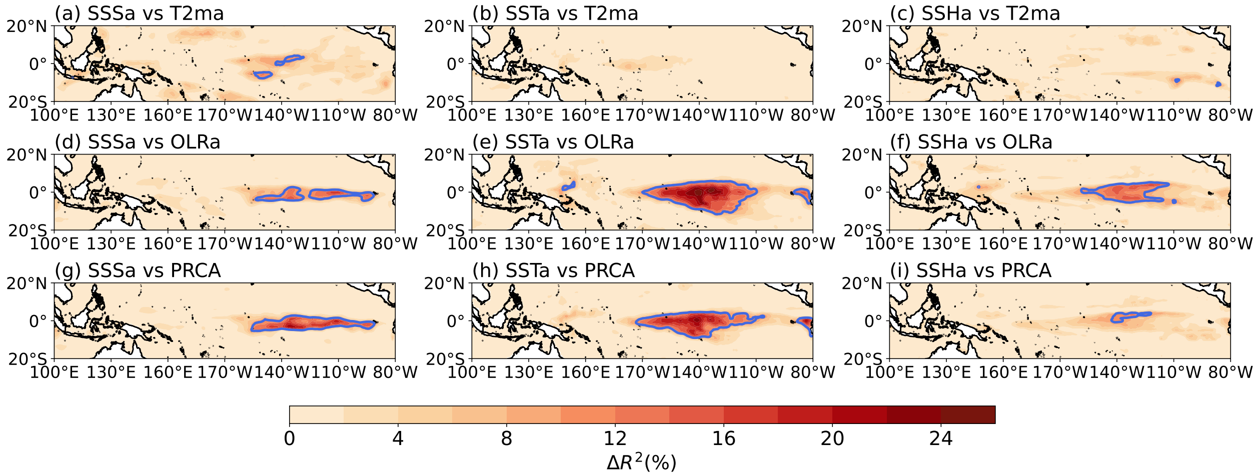

Fig. 1 Difference in R^2 between the quadratic fitting model and linear regression model for modeling the correlations between local monthly anomalies. The blue line denotes that R^2 of quadratic fitting is 10 percent larger than linear regression

represents the Kalman gain matrix, which can be written as

$$K = \boldsymbol{B}\boldsymbol{H}^T\left(\boldsymbol{H}\boldsymbol{B}\boldsymbol{H}^T + \boldsymbol{R}\right)^{-1} \tag{2}$$

where $\boldsymbol{B}$ denotes the flow dependent background error covariance matrix, a statistical variance–covariance matrix estimated by ensemble members to characterize the correlation of variables among the model grid points. $\boldsymbol{R}$ is the observation error covariance matrix that can be derived from the observation error of instruments.

In order to analyze the increment of prior prediction induced by the observations from another earth component during one assimilation cycle, here we consider a two-variable one-dimensional field denoted as (x, y), with x and y represent oceanic and atmospheric components, respectively. To directly illustrate the adjustment and highlight the role of $\boldsymbol{B}$ in adjustment (Appendix B), we further simplify the update equation by assuming that there are only accurate observations of oceanic variable x_o available at the model grid points, so we have $\boldsymbol{X}^o = x_o$, $\boldsymbol{R} = 0$ and $\boldsymbol{H} = (1{,}0)$. Supposing that the prior model prediction is $\boldsymbol{X}^p = (x_p, y_p)^T$, the update equation finally turns to

$$\boldsymbol{X}^a = \boldsymbol{X}^p + \boldsymbol{B}\boldsymbol{H}^T\left(\boldsymbol{H}\boldsymbol{B}\boldsymbol{H}^T\right)^{-1}\delta X \tag{3}$$

where $\delta\boldsymbol{X} = \boldsymbol{X}^o - \boldsymbol{H}\boldsymbol{X}^p = x_o - x_p = \delta x$, so that the analyses of x and y are computed by

$$x_a = x_p + \delta x \tag{4}$$

$$y_a = y_p + \frac{\sigma_{yx}}{\sigma_x^2}\delta x \tag{5}$$

where σ_{yx} represents the covariance between y and x, σ_x^2 represents the standard deviation of the x. σ_{yx} and σ_x^2 are parts of $\boldsymbol{B}$. Equation (5) clearly shows that the oceanic observation innovation is projected to atmospheric component through a linear coefficient, which is the ratio of σ_{yx} and σ_x^2.

If we substitute the variance–covariance correlation with linear regression to describe the relationship between air–sea variables, then the analysis field will be expressed as

$$x_a = x_p + 1 \times \delta x + 0 \tag{6}$$

$$y_a = y_p + a \times \delta x + b \tag{7}$$

where 0 and b represent the bias, and the regression coefficient a is calculated by

$$a = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} = \frac{\frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N}}{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}} = \frac{\sigma_{xy}}{\sigma_x^2} \tag{8}$$

Therefore, the only difference between two assimilation strategies is the bias in update equation of y, that means SCDA based on the variance–covariance correlation is analogous to linear regression mathematically. These deductions are in agreement with previous works (Anderson 2003; Zhang et al. 2007). Equations (5) and (8) collectively demonstrate that through conventional SCDA, observations can only linearly impact the adjustment of state from different components.

Correlation between variables from different components

Tropical Pacific is a region characterized by intense air–sea interaction. To investigate the importance of nonlinear part in the relationships between different earth components, we evaluate the difference of determinable coefficient (R^2) between the second-order and first-order Taylor expansion of the function $y = f(x)$ over

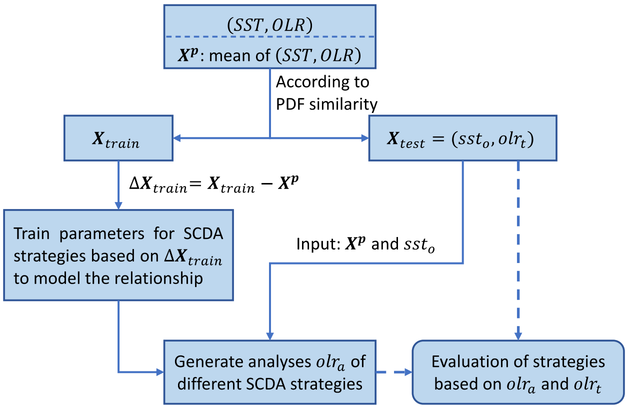
(SST, OLR)
$\boldsymbol{X}^{\boldsymbol{p}}$: mean of (SST, OLR)
According to PDF similarity
$\boldsymbol{X}_{train}$
$\boldsymbol{X}_{test} = (sst_o, olr_t)$
$\Delta\boldsymbol{X}_{train} = \boldsymbol{X}_{train} - \boldsymbol{X}^{\boldsymbol{p}}$
Train parameters for SCDA strategies based on $\Delta\boldsymbol{X}_{train}$ to model the relationship
Input: $\boldsymbol{X}^{\boldsymbol{p}}$ and sst_o
Generate analyses olr_a of different SCDA strategies
Evaluation of strategies based on olr_a and olr_t

Fig. 2 Nonlinear assimilation experimental design

$(20°\text{S} - 20°\text{N}, 100°\text{E} - 80°\text{W})$, represented by linear regression model and quadratic fitting model, respectively. R^2 is an important metric for quantifying the explanatory power of the model, with larger values indicating a more adaptive model to the task. The mathematical formulation is expressed as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{9}$$

where n is the number of the truth of atmospheric components y_i, $\widehat{y}_i$ is the estimates from fitting models and $\overline{y}$ is the mean of the truth. Decadal variability is removed from data by subtracting the monthly mean at different locations.

From Fig. 1, we observe that nonlinear correlations are prevalent among variables in the tropical Pacific. Utilizing the quadratic fitting model to characterize these relationships can enhance R^2 more than 10%, even larger than 24% in certain instances. It suggests that nonlinear component plays a crucial role and nonlinear approaches are more suitable to describe the relationships between variables from different earth components. However, the conventional SCDA only captures the linear correlation between variables to adjust the initial conditions for the coupled models, which conflicts with the realistic nonlinear relationships. SCDA may lead the predictions unreliable, along with the linearized model and measurement operator.

Although nonlinear correlation also exists within a single earth component, the fact that all variables are governed by the same dynamic equations helps mitigate the negative impact caused by the linear deficiency of assimilation algorithms. However, it remains challenging for coupled system, subject to different dynamic equations, to avoid the incongruity between variables and address the linear deficiency. These inferences suggest that introducing non-linearity into SCDA methods is necessary for generating more accurate analysis field and reliable predictions.

Experiment settings for strategy effectiveness evaluation

Multilayer perceptron (MLP)

MLP is a fully connected feedforward artificial neural network, comprising an input layer, at least one hidden layer and a output layer (Subasi 2020). The number of hidden layers and neurons is task-dependent. Each neuron introduces nonlinearity through a nonlinear activation function, enabling the MLP to approximate any complex functions. The weights help establish connections between neurons in adjacent layers. The network gradually converges to the optimal solution by minimizing the loss function. MLP has exhibited remarkable performance in various applications, including image recognition and pattern recognition. Flexibility, a notable strength of MLP, enables the network to tackle a variety of tasks (Hornik et al. 1989). MLP remains effective even when faced with a single input feature, underscoring its ability to perform well in low-dimensional input settings (Taud and Mas 2018).

Experiment design

Under assumptions in Sect. "Linear characteristic of coupled data assimilation", three SCDA strategies will be tested to handle relations between SST and OLR with different degrees of nonlinearity: quadratic fitting, a data-informed nonlinear polynomial fitting model; MLP, an effective data-driven and adaptive nonlinear machine learning approach; and EnKF, representing variance–covariance correlation and serving as a baseline for comparison.

Several temporally corresponding pairs of (sst_o, olr_t) points will be randomly selected from SST and OLR data sets in Sect. "Datasets" as test data set $\boldsymbol{X}_{test}$, with sst_o serves as the x_o in Sect. "Linear characteristic of coupled data assimilation" and olr_t as the atmospheric truth to assess the analyses olr_a generated by different SCDA strategies. Similarity is crucial for avoiding the deviation from realistic situation and effectively evaluating strategies, so the number of pairs depends on the value of JSD between the distributions of selected points and the real situation (estimated by the entire data set). The remaining part serves as the training data set $\boldsymbol{X}_{train}$. The mean of the entire data set serves as $\boldsymbol{X}^p$ and $\boldsymbol{B}$ is estimated by anomalies $\Delta\boldsymbol{X}_{train}$, where Δ denotes $\boldsymbol{X}_{train}$ minus $\boldsymbol{X}^p$. $\Delta\boldsymbol{X}_{train}$ will be utilized to train parameters for strategies to model the relationship (Fig. 2).

Table 1 Statistical average evaluation metrics of different strategies for relations with different complexity

		Near-linear	Weak nonlinear	Strong nonlinear
RMSE (W/m^2)	EnKF (Var-Cov)	14.614	15.227	9.052
	Quadratic fitting	14.514	13.543	6.630
	MLP	14.294	13.076	5.897
	Improvement	**− 3%**	**− 14%**	**− 35%**
MAE (W/m^2)	EnKF (Var-Cov)	11.639	10.866	6.144
	Quadratic fitting	11.525	9.436	4.550
	MLP	11.349	8.998	3.900
	Improvement	**− 2%**	**− 17%**	**− 37%**
R^2	EnKF (Var-Cov)	0.25	0.32	0.29
	Quadratic fitting	0.26	0.47	0.61
	MLP	0.28	0.51	0.68
	Improvement	**+3%**	**+19%**	**+39%**
Corr	EnKF (Var-Cov)	0.45	0.58	0.55
	Quadratic fitting	0.50	0.70	0.79
	MLP	0.52	0.73	0.83
	Improvement	**+7%**	**+15%**	**+28%**

The correlation coefficients are significant at the 5% significant level and the "improvement" rows in the table represent the improvement of the MLP compared to EnKF

To facilitate training, the data for MLP is initially normalized within the range of (− 1,1). The rectified linear unit (ReLU) is chosen as the activation function to process the input nonlinearly. The Mean Square Error (MSE) is employed as the loss function and the Adaptive Moment Estimation (Adam) algorithm serves as weight updating scheme. The learning rate, as well as the depth and width of MLP are determined empirically. Considering the data volume, K-fold cross-validation is introduced to determine the optimal validation data set and trained model.

Evaluation metrics

To systematically measure the performance of different assimilation strategies, six evaluation metrics are introduced. Besides R^2, the root-mean-square error (RMSE) and mean absolute error (MAE) serve to quantify the precision of the assimilation model. Compared to MAE, RMSE amplifies the contribution of larger errors to comprehensive performance of the assimilation model, which could help evaluate the performance of model in the extreme events. Pearson correlation coefficient (Corr) is used to evaluate the degree of coordinated variation of the analyses and the truth olr_t.

Considerable emphasis should also be placed on whether the distribution of the analysis field deviates from the truth. The Kullback–Leribler divergence (D_{KL}), also known as relative entropy, is commonly used to evaluate the disparity between two distributions. D_{KL} is defined as

$$D_{KL}(P||Q) = \sum_i P(i)\log\left(\frac{P(i)}{Q(i)}\right) \tag{10}$$

where P is the baseline distribution and Q is the sample distribution. D_{KL} is a non-negative asymmetric metric, signifying that $D_{KL}(P||Q) \neq D_{KL}(Q||P)$. The lower the values of D_{KL}, the smaller the disparity between Q and P.

The Jensen–Shannon divergence (JSD) is a more comprehensive measure of distribution similarity, which inherits the capabilities of D_{KL} and addresses its asymmetry deficiency simultaneously. The formula is as follows:

$$JSD(P,Q) = \frac{1}{2}D_{KL}(P||M) + \frac{1}{2}D_{KL}(Q||M) \tag{11}$$

where M is the average of P and Q. The range of JSD is 0–1. Here we define that the difference between P and Q is negligible when $JSD \leq 0.01$.

Performance of different SCDA strategies

We first evaluate the effectiveness of strategies in near-linear, weak and strong nonlinear relations at local grid points within the tropical pacific, where the data assimilation is practiced on. 100 points are selected to reflect

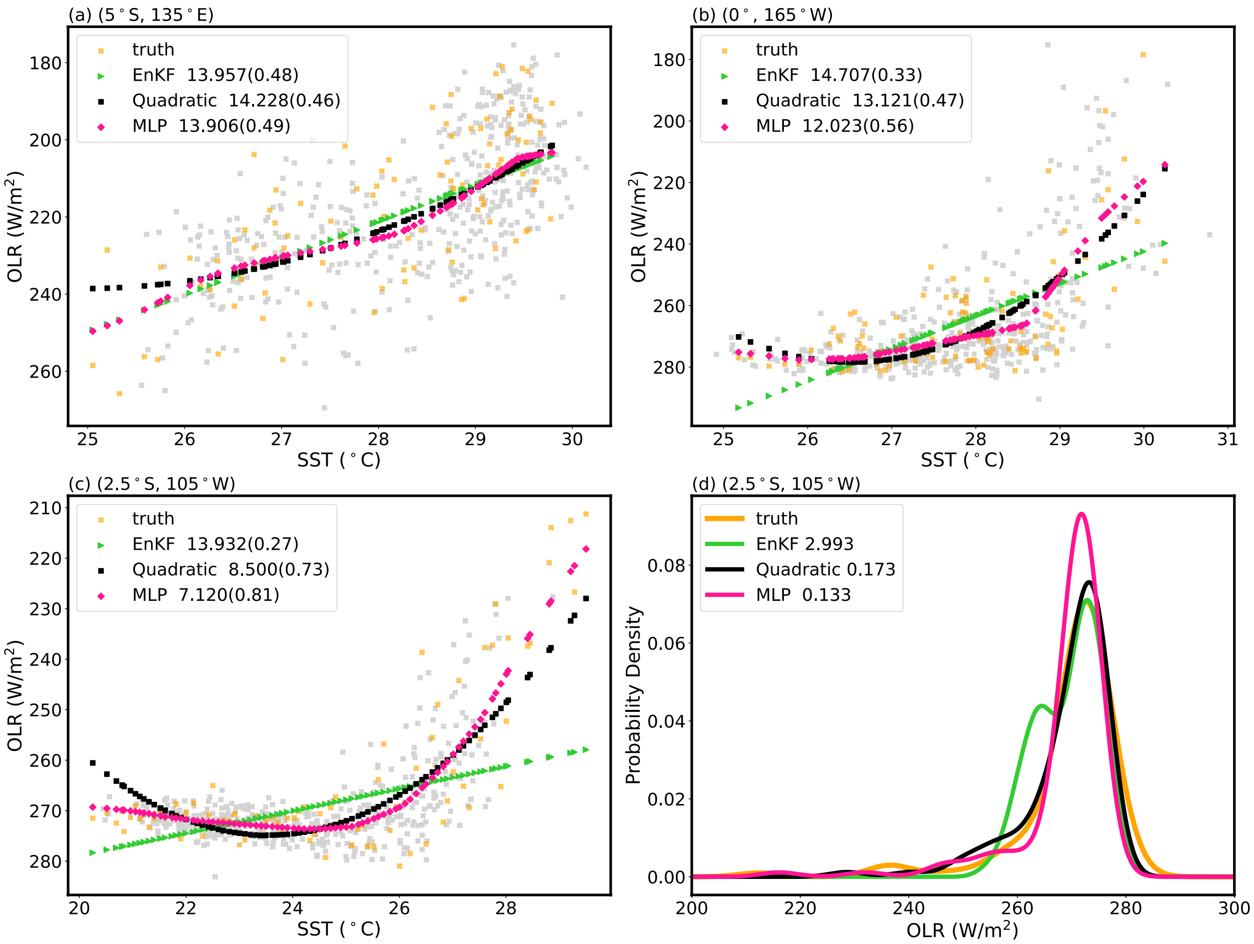

Fig. 3 Analyses produced by EnKF, quadratic fitting and MLP strategies during one assimilation cycle for relations between SST and OLR: **a** near-linear relation at (5°S, 135°E); **b** weak nonlinear relation at (0°, 165°W); **c** strong nonlinear relation at (2.5°S, 105°W); **d** analysis distributions for the strong nonlinear relation. In (**a**–**c**), the orange points denote the truth; the gray points represent the training data set and the values in the legend are *RMSE* (R^2). In (**d**), the values in the legend are D_{KL}, and a smaller value indicates a higher degree of similarity between the analysis distribution and the truth. To show clearly the performance of different strategies, points in (**c**) are selected evenly and the results depended on PDF are shown in Fig. 8c

the realistic situations for these examples according to the standard of *JSD* (Fig. 6), and 20 repeated experiments are conducted at different grid points for each condition (Table 1). Consistent with the mathematical deduction, Fig. 3 shows that the analysis of EnKF is a result of linear mapping. Figure 3a demonstrates that variance–covariance correlation (EnKF) may yield more accurate state estimation than the unsuitable data-informed quadratic fitting strategy under near-linear situation, though the latter shows statistically superiority (Table 1). When faced with nonlinear relations (Fig. 3b–d), linear strategy (i.e., EnKF) sacrifices information, especially in dealing with rare extreme events (OLR smaller than $240 W/m^2$). Besides modeling the ordinary events accurately, MLP strategy is more reliable in predicting extreme or out-of-sample events, which other strategies are prone to underestimate. However, both linear and data-informed nonlinear strategies could generate state estimation that deviate from the truth to different extent, particularly for “low probability, high impact” events. The advantage of flexibility enables data-driven MLP strategy adaptive to different relations, achieving statistically superior evaluation metrics (Table 1). The correlation coefficeinets in Table 1 also indicate that the linear strategy has difficulty capturing the evolving character of complex relationships. Figure 3d shows that the bimodal analysis of

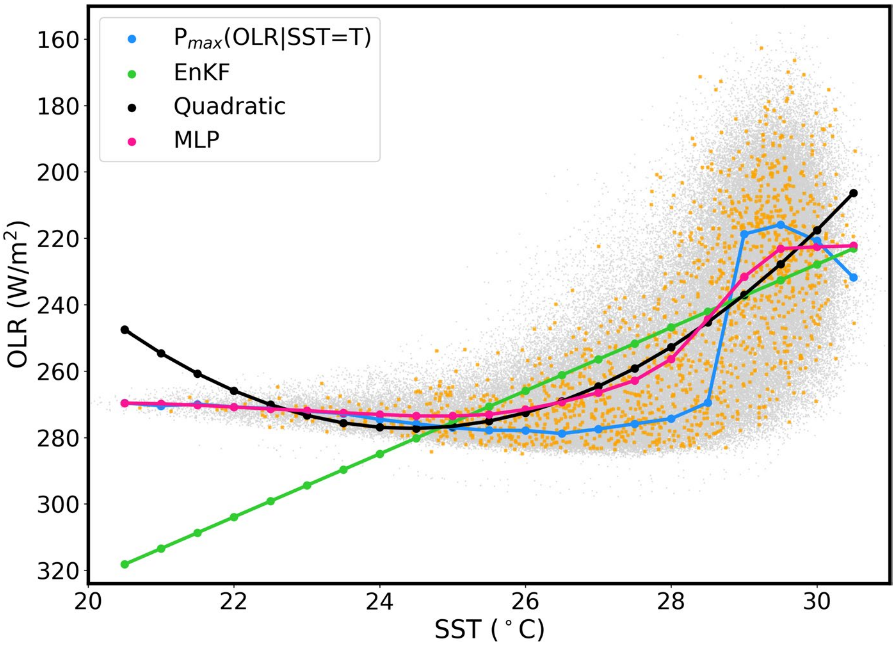

Fig. 4 Evolution of OLR with SST generated by different SCDA models at (5°S—5°N, 130°E—100°W). The blue points denote the OLR that has the maximum probability to happen for a given SST in a real situation and the interval between points is 0.5 °C. It shows that as SST increases, the value of OLR first increases ($SST \leq 25$ °C), remains relatively constant ($25\,°C \leq SST \leq 28\,°C$) and finally exhibits a curve relation with SST ($28\,°C \leq SST$). The other color points represent the analyses of OLR produced by different strategies for the given SST. The functions of gray and orange points remain the same as before

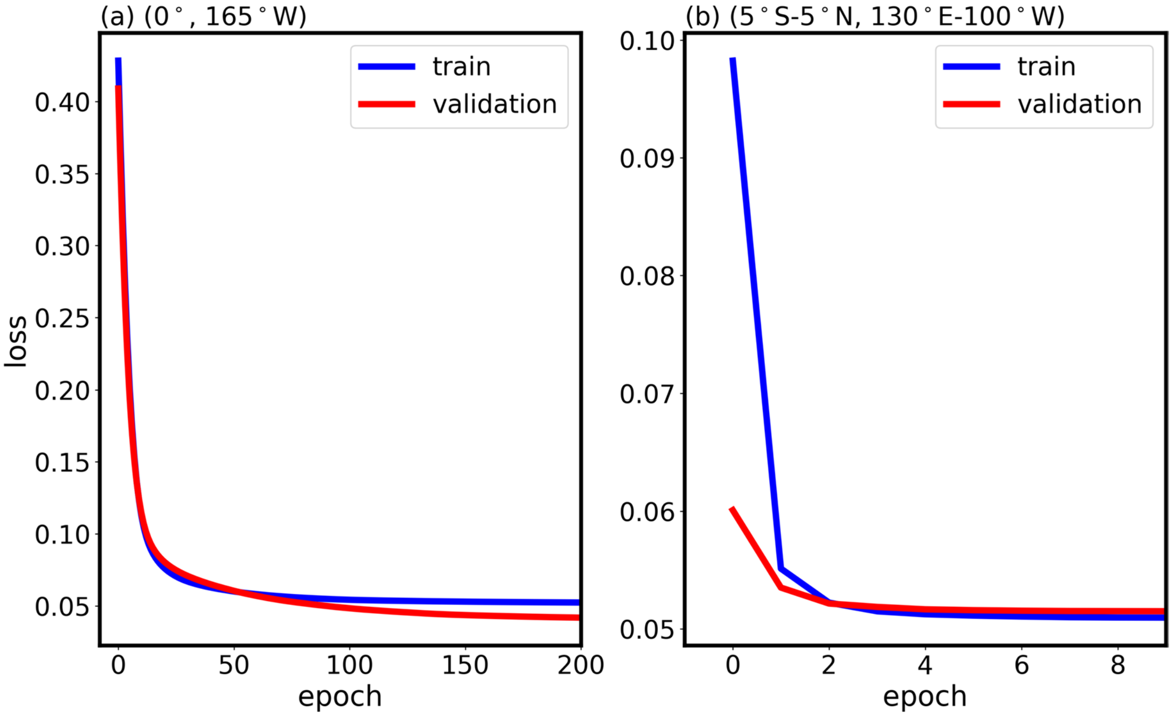

Fig. 5 Evolution of the loss function with epochs for the training and validation data set of relations between ΔSST and ΔOLR at (**a**) (0°, 165°W); (**b**) (5°S—5°N, 130°E—100°W)

EnKF is inconsistent with the unimodal truth, because the variance–covariance correlation is analogous to linear mapping, determining that EnKF will yield a wrong distribution consistent with SST but not OLR (Fig. 6). Nonlinear strategies all successfully reproduce the unimodal distribution and gain smaller D_{KL}.

Further comparison is made based on a more strongly nonlinear relationship between SST and OLR at $(5°\mathrm{S}{-}5°\mathrm{N}, 130°\mathrm{E}{-}100°\mathrm{W})$. 1500 points are selected for this example to validate the trained assimilation models (Fig. 7). The SCDA strategies based on least-squares criterion are not only required to reduce RMSE, but closely represent the true evolving relationship between variables. Here we use the evolution of OLR with SST based on joint probability distribution to evaluate the three trained models. The analysis produced by EnKF shows a large margin of error, even completely away from the range of OLR when SST is below 25 °C (Fig. 4). Other two nonlinear strategies generate more accurate state estimations without significant deviation. The data-informed quadratic fitting strategy is prone to overestimates the value of OLR when SST is smaller than 22 °C, and fails to reproduce the evolution of OLR when SST is larger than 28 °C. However, the data-driven MLP strategy effectively captures the various evolving character within this complex correlation (Lau et al. 1997; Jiang and Zhu 2020). The evolution reproduced by MLP closely matches the truth when SST is smaller than 25 °C, whereas other strategies exhibit substantially increased margins of error.

These inferences imply that the linear variance–covariance correlation is not suitable for modeling ubiquitous nonlinear relationships, resulting in substantial errors in the analyses and predictions. Introducing nonlinear strategies will remedy the linear limitation of conventional SCDA and improve the prediction skills for coupled models. Utilizing nonlinear fitting strategies to improve the state estimation proves to be more computationally expensive (Appendix C). In contrast, the ML strategy circumvents the need for constructing $\boldsymbol{B}$ explicitly, emerging as a promising approach for implementing SCDA in coupled models. Figure 5 also demonstrates that the loss function will converge stably and quickly as data volume grows, contributing to an improvement in computational efficiency of SCDA.

Conclusions and discussion

This study aims to clarify that the conventional SCDA based on the linear variance–covariance correlation faces limitations in addressing complex relations within coupled systems. The simplified two-variable one-dimensional nonlinear assimilation experiments based on SST and OLR are conducted in the tropical Pacific, a region characterized by intense air–sea interaction (Wang et al. 2017a). Experimental results indicate that the conventional SCDA strategy (i.e. ensemble-based assimilation method) is suitable for near-linear situations, but fails to represent nonlinear relationships. Given the universal nonlinear relationships in the real world, it is necessary to develop nonlinear SCDA strategies. Instead, the data-driven advantage enables ML strategy, represented by MLP here, to overcome the linear or Gaussian limitations of conventional strategies and adapt to relations with various complexity. This strategy also achieves comprehensively improved analysis quality than linear and data-informed nonlinear strategies, especially for regions with strongly nonlinear interaction between earth components. The superior results of evaluation metrics and the longer tail of analysis distributions collectively illustrate the significant potential effectiveness of ML in generating more accurate analysis field and enhancing the prediction skills of small probability events (Frame et al. 2022). In addition, it circumvents the explicit construction of background matrix and subsequent costly matrix operations, presenting a cost-effective approach for implementing SCDA. The augment of the data volume and input features could further enhance the computational efficiency of SCDA based on ML strategy, which can be achieved by increasing the ensemble size and integrating heterogeneous data from different sources. However, the limitation of ML in handling imbalanced data set may hinder the further improvement, solutions like introducing physical mechanism into ML strategy become imperative (Xie et al. 2021).

Appendix A

Distribution similarity and loss evolution

See Figs. 6, 7, 8, 9

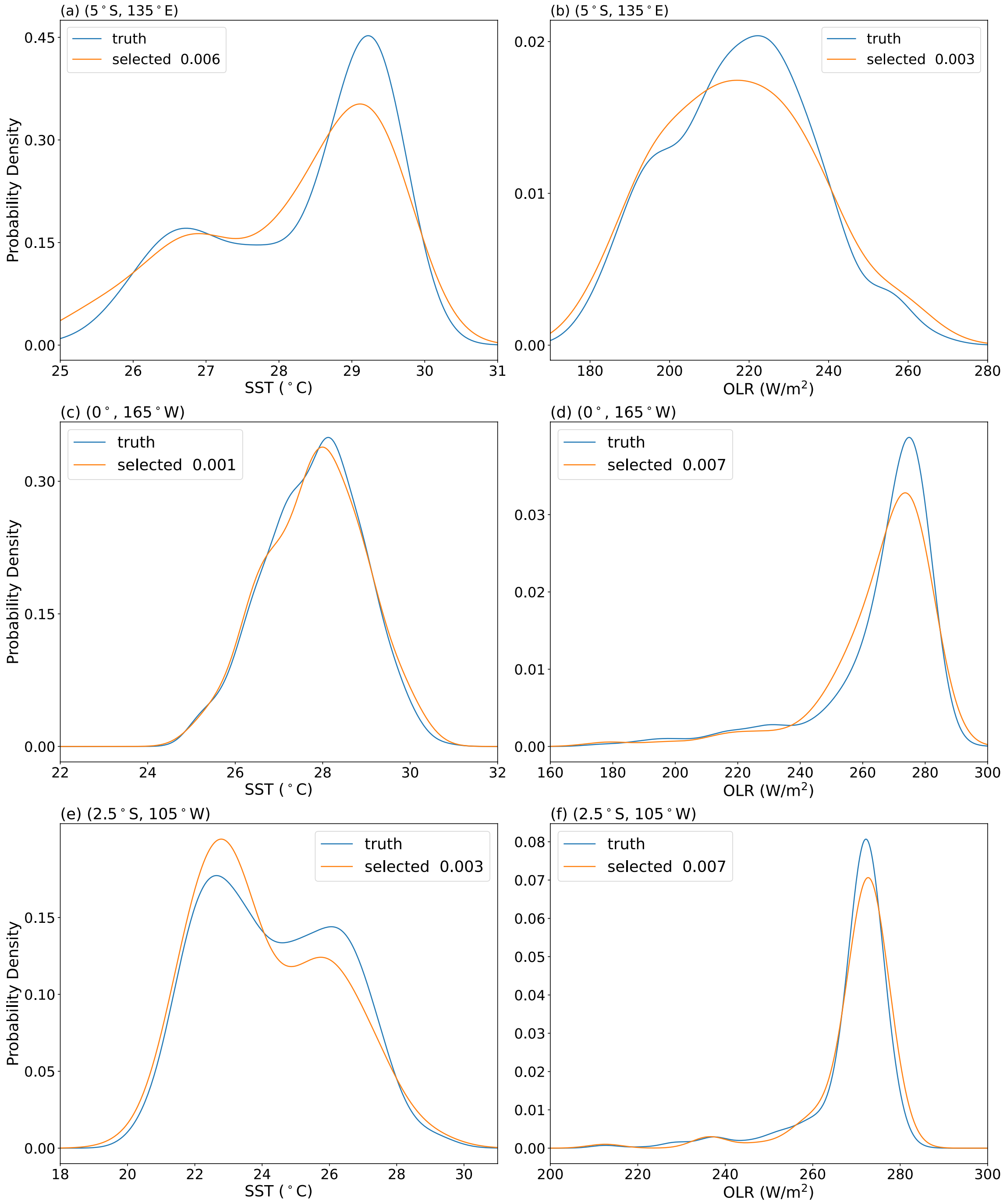

Fig. 6 *JSD* for examples of local relation between SST and OLR, when 100 points are selected. **a**, **b** at (5°S, 135°E); **c**, **d** at (5°S, 155°W); **e**, **f** at (2.5°S, 105°W). The values after the "selected" are *JSD*

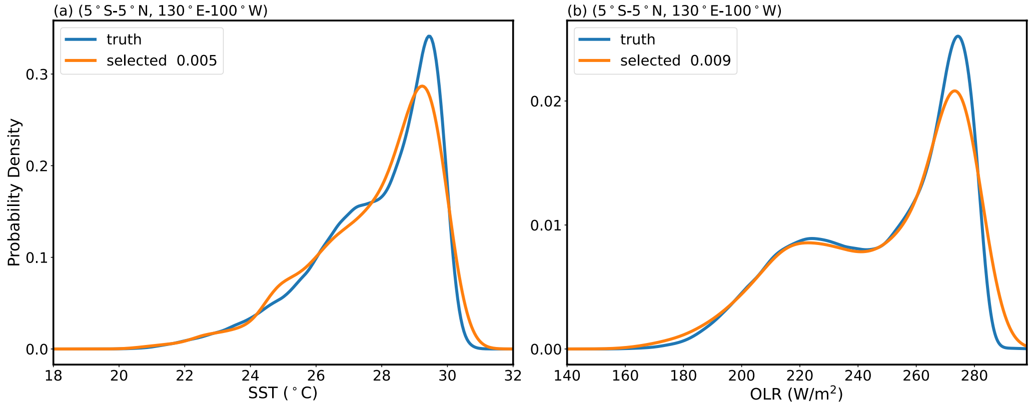

Fig. 7 Same as Figure A1 but for the relation between SST and OLR at (5°S—5°N, 130°E—100°W), when 1500 points are selected

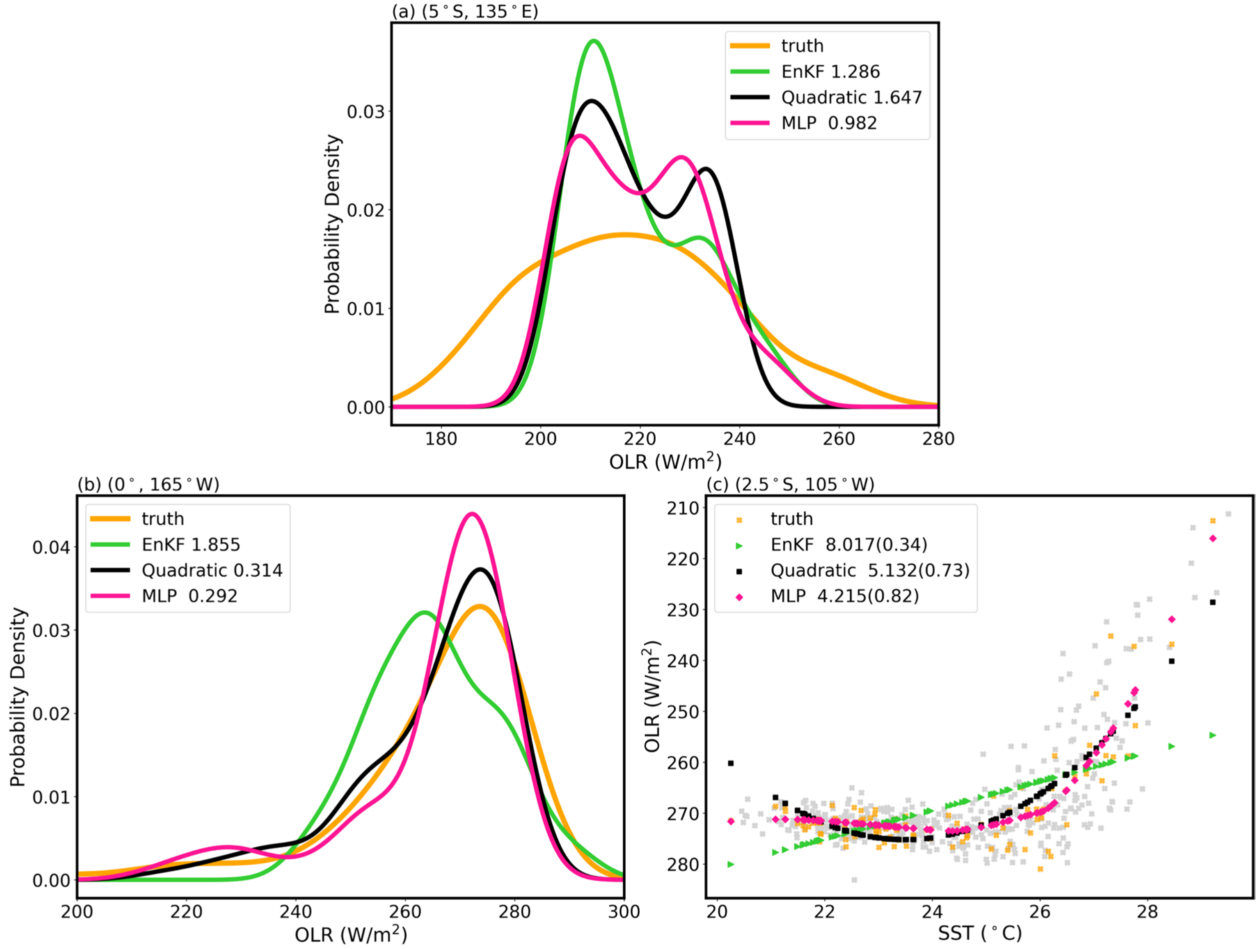

Fig. 8 Analysis distributions produced by EnKF, quadratic fitting and MLP strategies during one assimilation cycle for the examples of the relation between SST and OLR: **a** near-linear relation at (5°S, 135°E); **b** weak nonlinear relation at (0°, 165°W). In (**a**, **b**), the values in the legend are D_{KL}. **c** Analyses produced by different strategies during one assimilation cycle for strong nonlinear relation at (2.5°S, 105°W)

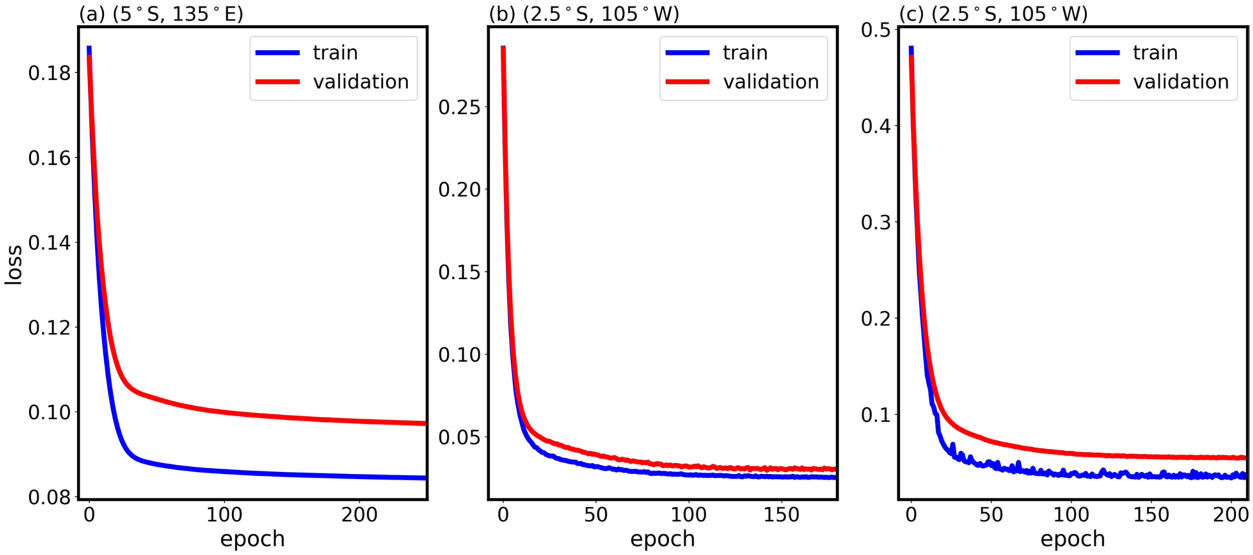

Fig. 9 Evolution of the loss function with epochs for the training and validation data set for relations between Δ*SST* and Δ*OLR* at (**a**) (5°S, 135°E); the PDF-dependent (**b**) and the evenly selected example (**c**) at (2.5°S, 105°W)

Appendix B

Influence of observation error covariance in adjustment

In real-world assimilation experiments, the observation error covariance $\boldsymbol{R}$ can be derived from the observation error of instruments and is usually non-zero. In this study, $\boldsymbol{R}$ is also a changeable positive variance, when only oceanic observation x_o is available. If we set $\boldsymbol{R} = \boldsymbol{c} * \sigma_x^2$, where $\boldsymbol{c}$ is a changeable positive constant and σ_x^2 is the background error variance of oceanic component. Then the update Eq. (1) turns to:

$$\boldsymbol{X}^a = \boldsymbol{X}^p + \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{yx} & \sigma_y^2 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} \left(\begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{yx} & \sigma_y^2 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} + \boldsymbol{c} * \sigma_x^2 \right)^{-1} \delta \boldsymbol{X} \tag{12}$$

and the analyses of x and y are described as:

$$x_a = x_p + \frac{1}{1+\boldsymbol{c}}(x_o - x_p) \tag{13}$$

$$y_a = y_p + \frac{\sigma_{yx}}{\sigma_x^2}\frac{1}{1+\boldsymbol{c}}(x_o - x_p) \tag{14}$$

Equations (13) and (14) clearly shows that the oceanic observation innovation is projected to atmospheric component through a linear coefficient. When σ_x^2 is fixed, the value of $\boldsymbol{c}$ will influence the analyses generated by EnKF. To further clarify the influence of $\boldsymbol{R}$, Fig.10 provides an example based on SST and OLR at a local grid point. It clearly shows that when $0 \le \boldsymbol{c} < 1 (0 \le \boldsymbol{R} < \sigma_x^2)$, the SCDA model weights observation x_o more than prior prediction x_p and when $\boldsymbol{c} = 0$ ($\boldsymbol{R} = 0$), observation is completely trusted to modify the prior predictions $\boldsymbol{X}^p$. When $1 \le \boldsymbol{c}(\sigma_x^2 \le \boldsymbol{R})$, x_p plays more important role in assimilation, analyses $\boldsymbol{X}^a$ are gradually close to $\boldsymbol{X}^p$ as $\boldsymbol{c}$ increases. When $\boldsymbol{c} = \infty (\boldsymbol{R} = \infty)$, analyses $\boldsymbol{X}^a$ are finally equal to $\boldsymbol{X}^p$. Although $\boldsymbol{R}$ derived from instruments will influence the analyses, it will not change the linear characteristic of conventional SCDA strategies. However, the variance–covariance relationship (σ_{yx}/σ_x^2) between cross-sphere variables estimated by ensemble members can be replaced by a nonlinear form to supply nonlinearity to SCDA strategies.

See Fig. 10.

Although $\boldsymbol{R}$ derived from instruments will influence the analyses, it will not change the linear characteristic of conventional SCDA strategies. However, the variance–covariance relationship (σ_{yx}/σ_x^2) between cross-sphere variables estimated by ensemble members can be replaced by a nonlinear form to supply nonlinearity to SCDA strategies.

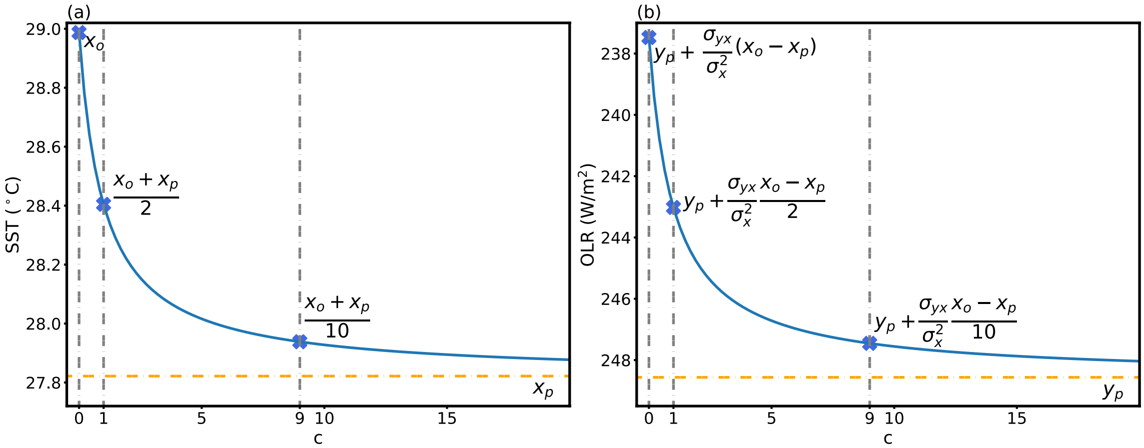

Fig. 10 Evolution of the analyses of (**a**) SST and (**b**) OLR with ***c*** at a local grid point, when only observation of SST x_o is available. The orange dashed lines denote the prior predictions. The texts represent the values of analyses (blue crosses) for the given ***c***

Appendix C

Construction of nonlinear background matrix

To enhance the accuracy of state estimation, we try to introduce nonlinearity into SCDA by constructing a nonlinear ***B*** based on quadratic fitting. The quadratic fitting offers the benefits of simplicity, nonlinearity, and a higher level of accuracy. Then the analysis formulations of x and y are:

$$x_a = x_p + \delta x \tag{15}$$

$$y_a = y_p + a\delta x^2 + b\delta x + c \tag{16}$$

where $\delta x = x_o - x_p$, the corresponding tangent linear format can be written as:

$$\delta x_a = \delta x_o \tag{17}$$

$$\delta y_a = 2a\left(x_o - x_p\right)\delta x_o + b\delta x_o \tag{18}$$

then the reconstructed nonlinear ***B*** can be obtained through inverse deduction:

$$B = \begin{pmatrix} 1 & 0 \\ 2a\left(x_o - x_p\right) + b & 0 \end{pmatrix} \tag{19}$$

And compared to the original one, the analysis equation has one more constant vector $\boldsymbol{C} = (0, c)^T$:

$$\boldsymbol{X^a} = \boldsymbol{X^p} + \boldsymbol{BH^T}\left(\boldsymbol{HBH^T}\right)^{-1}\delta X + C \tag{20}$$

This simplest scenario demonstrates that the nonlinear ***B*** in Eq. (19) is no longer a symmetrical constant matrix and the elements of ***B*** are functions dependent on observations. However, as the degree of polynomial fitting, variable and observation increase, along with segmented partitioning, the rapid growing complexity of ***B*** will result in a surge of computational cost. The adaptability of this matrix requires further examination. Therefore, more practical nonlinear strategies are needed for SCDA.

Abbreviations

CDA	Coupled data assimilation
SCDA	Strongly coupled data assimilation
WCDA	Weakly coupled data assimilation
CCEC	Coupled cross background-error covariance
LACC	Leading averaged coupled covariance
SST	Sea surface temperature
SSS	Sea surface salinity
SSH	Sea surface height
OLR	Outgoing longwave radiation
PRC	Precipitation rate
T2m	The air temperature at 2 m
EnKF	Ensemble Kalman filter
ML	Machine learning
MLP	Multilayer perceptron
R^2	Determinable coefficient
RMSE	Root-mean-square error
MAE	Mean absolute error
Corr	Pearson correlation coefficient
D_{KL}	Kullback–Leribler divergence
JSD	Jensen–Shannon divergence

Acknowledgements

This work was supported by the National Key R&D Program of China (Grant No. 2023YFF0805202), the National Natural Science Foundation of China (Grant No. 42175045), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDB42000000).

Author contributions

Ziying Xuan performed the data analysis and wrote the original draft. Fei Zheng provided the funding acquisition, and guided the revision of the paper. All authors read and approved the final manuscript.

Funding

This work was supported by the National Key R&D Program of China (Grant No. 2023YFF0805202), the National Natural Science Foundation of China (Grant No. 42175045), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDB42000000).

Availability of data and materials

The data sets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 8 April 2024 Accepted: 11 July 2024
Published online: 18 September 2024

References

Adler RF, Huffman GJ, Chang A et al (2003) The version-2 global precipitation climatology project (GPCP) monthly precipitation analysis (1979–present). J Hydrometeorol 4:1147–1167. https://doi.org/10.1175/1525-7541(2003)004%3c1147:TVGPCP%3e2.0.CO;2

Anderson JL (2003) A local least squares framework for ensemble filtering. Mon Weather Rev 131:634–642. https://doi.org/10.1175/1520-0493(2003)131%3c0634:ALLSFF%3e2.0.CO;2

Arcucci R, Zhu J, Hu S, Guo Y-K (2021) Deep data assimilation: integrating deep learning with data assimilation. Appl Sci 11:1114. https://doi.org/10.3390/app11031114

Boer GJ, Smith DM, Cassou C et al (2016) The decadal climate prediction project (DCPP) contribution to CMIP6. Geosci Model Dev 9:3751–3777. https://doi.org/10.5194/gmd-9-3751-2016

Brajard J, Carrassi A, Bocquet M, Bertino L (2020) Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: a case study with the Lorenz 96 model. J Comput Sci 44:101171. https://doi.org/10.1016/j.jocs.2020.101171

Carton JA, Chepurin GA, Chen L (2018) SODA3: a new ocean climate reanalysis. J Clim 31:6967–6983. https://doi.org/10.1175/JCLI-D-18-0149.1

Evensen G, Vossepoel FC, van Leeuwen PJ (2022) Data assimilation fundamentals: a unified formulation of the state and parameter estimation problem. Springer Nature, Cham

Feng J, Wang X, Poterjoy J (2020) A comparison of two local moment-matching nonlinear filters: local particle filter (LPF) and local nonlinear ensemble transform filter (LNETF). Mon Weather Rev 148:4377–4395. https://doi.org/10.1175/MWR-D-19-0368.1

Frame JM, Kratzert F, Klotz D et al (2022) Deep learning rainfall–runoff predictions of extreme events. Hydrol Earth Syst Sci 26:3377–3392. https://doi.org/10.5194/hess-26-3377-2022

Fujii Y, Ishibashi T, Yasuda T et al (2021) Improvements in tropical precipitation and sea surface air temperature fields in a coupled atmosphere–ocean data assimilation system. Q J R Meteorol Soc 147:1317–1343. https://doi.org/10.1002/qj.3973

Gouretski V, Reseghetti F (2010) On depth and temperature biases in bathythermograph data: development of a new correction scheme based on analysis of a global ocean database. Deep Sea Res Part I 57:812–833. https://doi.org/10.1016/j.dsr.2010.03.011

Han G, Wu X, Zhang S et al (2013) Error covariance estimation for coupled data assimilation using a lorenz atmosphere and a simple pycnocline ocean model. J Clim 26:10218–10231. https://doi.org/10.1175/JCLI-D-13-00236.1

He Y, Wang B, Liu M et al (2017) Reduction of initial shock in decadal predictions using a new initialization strategy. Geophys Res Lett 44:8538–8547. https://doi.org/10.1002/2017GL074028

He Y, Wang B, Huang W et al (2020) A new DRP-4DVar-based coupled data assimilation system for decadal predictions using a fast online localization technique. Clim Dyn 54:3541–3559. https://doi.org/10.1007/s00382-020-05190-w

Hersbach H, Bell B, Berrisford P et al (2020) The ERA5 global reanalysis. Q J R Meteorol Soc 146:1999–2049. https://doi.org/10.1002/qj.3803

Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators. Neural Netw 2:359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Huang L, Leng H, Li X et al (2021) A data-driven method for hybrid data assimilation with multilayer perceptron. Big Data Res 23:100179. https://doi.org/10.1016/j.bdr.2020.100179

Jiang N, Zhu C (2020) Tropical Pacific cold tongue mode triggered by enhanced warm pool convection due to global warming. Environ Res Lett 15:054015. https://doi.org/10.1088/1748-9326/ab7d5e

Kalnay E, Sluka T, Yoshida T et al (2023) Review article: towards strongly coupled ensemble data assimilation with additional improvements from machine learning. Nonlinear Process Geophys 30:217–236. https://doi.org/10.5194/npg-30-217-2023

Lau K-M, Wu H-T, Bony S (1997) The role of large-scale atmospheric circulation in the relationship between tropical convection and sea surface temperature. J Clim 10:381–392. https://doi.org/10.1175/1520-0442(1997)010%3c0381:TROLSA%3e2.0.CO;2

Liebmann B, Smith CA (1996) Description of a complete (interpolated) outgoing longwave radiation dataset. Bull Am Meteor Soc 77:1275–1277

Liu Z, Wu S, Zhang S et al (2013) Ensemble data assimilation in a simple coupled climate model: the role of ocean-atmosphere interaction. Adv Atmos Sci 30:1235–1248. https://doi.org/10.1007/s00376-013-2268-z

Lu F, Liu Z, Zhang S, Liu Y (2015) Strongly coupled data assimilation using leading averaged coupled covariance (LACC). Part I: simple model study. Mon Weather Rev 143:3823–3837. https://doi.org/10.1175/MWR-D-14-00322.1

Park SK, Lim S, Zupanski M (2015) Structure of forecast error covariance in coupled atmosphere–chemistry data assimilation. Geosci Model Dev 8:1315–1320. https://doi.org/10.5194/gmd-8-1315-2015

Penny SG, Akella S, Buehner M et al (2017) Coupled data assimilation for integrated earth system analysis and prediction: goals, challenges, and recommendations. World Meteorological Organization, WWRP 2017-3, 50. https://library.wmo.int/doc_num.php?explnum_id=10830. Accessed 21 Jun 2023

Penny SG, Bach E, Bhargava K et al (2019) Strongly coupled data assimilation in multiscale media: experiments using a quasi-geostrophic coupled model. J Adv Model Earth Syst 11:1803–1829. https://doi.org/10.1029/2019MS001652

Poterjoy J (2016) A localized particle filter for high-dimensional nonlinear systems. Mon Weather Rev 144:59–76. https://doi.org/10.1175/MWR-D-15-0163.1

Raymond C, Horton RM, Zscheischler J et al (2020) Understanding and managing connected extreme events. Nat Clim Chang 10:611–621. https://doi.org/10.1038/s41558-020-0790-4

Rayner NA, Parker DE, Horton EB et al (2003) Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. J Geophys Res: Atmos. https://doi.org/10.1029/2002JD002670

Ruckstuhl Y, Janjić T, Rasp S (2021) Training a convolutional neural network to conserve mass in data assimilation. Nonlinear Process Geophys 28:111–119. https://doi.org/10.5194/npg-28-111-2021

Sakov P, Sandery PA (2015) Comparison of EnOI and EnKF regional ocean reanalysis systems. Ocean Model 89:45–60. https://doi.org/10.1016/j.ocemod.2015.02.003

Sarker IH (2021) Machine learning: algorithms, real-world applications and research directions. SN Comput Sci 2:160. https://doi.org/10.1007/s42979-021-00592-x

Skauvold J, Eidsvik J, van Leeuwen PJ, Amezcua J (2019) A revised implicit equal-weights particle filter. Q J R Meteorol Soc 145:1490–1502. https://doi.org/10.1002/qj.3506

Sluka TC, Penny SG, Kalnay E, Miyoshi T (2016) Assimilating atmospheric observations into the ocean using strongly coupled ensemble data assimilation. Geophys Res Lett 43:752–759. https://doi.org/10.1002/2015GL067238

Smith PJ, Lawless AS, Nichols NK (2018) Treating sample covariances for use in strongly coupled atmosphere-ocean data assimilation. Geophys Res Lett 45:445–454. https://doi.org/10.1002/2017GL075534

Subasi A (2020) Machine learning techniques. In: Subasi A (ed) Practical machine learning for data analysis using python. Academic Press, Cambridge, pp 91–202

Sun J, Liu Z, Lu F et al (2020) Strongly coupled data assimilation using leading averaged coupled covariance (LACC). Part III: assimilation of real world reanalysis. Mon Weather Rev 148:2351–2364. https://doi.org/10.1175/MWR-D-19-0304.1

Taud H, Mas JF (2018) Multilayer perceptron (MLP). In: Camacho Olmedo MT, Paegelow M, Mas J-F, Escobar F (eds) Geomatic approaches for modeling land change scenarios. Springer International Publishing, Cham, pp 451–455

Tödter J, Ahrens B (2015) A second-order exact ensemble square root filter for nonlinear data assimilation. Mon Weather Rev 143:1347–1367. https://doi.org/10.1175/MWR-D-14-00108.1

Tondeur M, Carrassi A, Vannitsem S, Bocquet M (2020) On temporal scale separation in coupled data assimilation with the ensemble kalman filter. J Stat Phys 179:1161–1185. https://doi.org/10.1007/s10955-020-02525-z

Wang C, Deser C, Yu J-Y et al (2017a) El niño and southern oscillation (ENSO): a review. In: Glynn PW, Manzello DP, Enochs IC (eds) Coral reefs of the eastern tropical pacific: persistence and loss in a dynamic environment. Springer, Dordrecht, pp 85–106

Wang X, Jiang D, Lang X (2017b) Future extreme climate changes linked to global warming intensity. Sci Bull 62:1673–1680. https://doi.org/10.1016/j.scib.2017.11.004

Xie K, Liu P, Zhang J et al (2021) Physics-guided deep learning for rainfall-runoff modeling by considering extreme events and monotonic relationships. J Hydrol 603:127043. https://doi.org/10.1016/j.jhydrol.2021.127043

Xu L, Chen N, Chen Z et al (2021) Spatiotemporal forecasting in earth system science: methods, uncertainties, predictability and future directions. Earth Sci Rev 222:103828. https://doi.org/10.1016/j.earscirev.2021.103828

Yoshida T, Kalnay E (2018) Correlation-cutoff method for covariance localization in strongly coupled data assimilation. Mon Weather Rev 146:2881–2889. https://doi.org/10.1175/MWR-D-17-0365.1

Yu X, Zhang S, Li J et al (2019) A Multi-timescale EnOI-like high-efficiency approximate filter for coupled model data assimilation. J Adv Model Earth Syst 11:45–63. https://doi.org/10.1029/2018MS001504

Zhang S (2011) A study of impacts of coupled model initial shocks and state-parameter optimization on climate predictions using a simple pycnocline prediction model. J Clim 24:6210–6226. https://doi.org/10.1175/JCLI-D-10-05003.1

Zhang M, Zhang F (2012) E4DVar: coupling an ensemble kalman filter with four-dimensional variational data assimilation in a limited-area weather prediction model. Mon Weather Rev 140:587–600. https://doi.org/10.1175/MWR-D-11-00023.1

Zhang S, Harrison MJ, Rosati A, Wittenberg A (2007) System design and evaluation of coupled ensemble data assimilation for global oceanic climate studies. Mon Weather Rev 135:3541–3564. https://doi.org/10.1175/MWR3466.1

Zhang S, Liu Z, Zhang X et al (2020) Coupled data assimilation and parameter estimation in coupled ocean–atmosphere models: a review. Clim Dyn 54:5127–5144. https://doi.org/10.1007/s00382-020-05275-6

Zheng F, Zhu J (2010) Coupled assimilation for an intermediated coupled ENSO prediction model. Ocean Dyn 60:1061–1073. https://doi.org/10.1007/s10236-010-0307-1

Zheng F, Liu J-P, Fang X-H et al (2022) The predictability of ocean environments that contributed to the 2020/21 extreme cold events in China: 2020/21 la niña and 2020 arctic sea ice loss. Adv Atmos Sci 39:658–672. https://doi.org/10.1007/s00376-021-1130-y

Zhou L, Zhang R-H (2023) A self-attention–based neural network for three-dimensional multivariate modeling and its skillful ENSO predictions. Sci Adv 9:2827. https://doi.org/10.1126/sciadv.adf2827

Zhu M, van Leeuwen PJ, Amezcua J (2016) Implicit equal-weights particle filter. Q J R Meteorol Soc 142:1904–1919. https://doi.org/10.1002/qj.2784

Publisher's Note