Journal of
Marine Science and Engineering

MDPI

Article

A Deep Neural Network-Ensemble Adjustment Kalman Filter and Its Application on Strongly Coupled Data Assimilation

Renxi Wang [1] and Zheqi Shen [1,2,3,*]

[1] College of Oceanography, Hohai University, Nanjing 210098, China; wrx_17330937822@163.com
[2] Key Laboratory of Marine Hazards Forecasting, Ministry of Natural Resources, Hohai University, Nanjing 210098, China
[3] Southern Laboratory of Ocean Science and Engineering (Zhuhai), Zhuhai 519082, China
* Correspondence: zqshen@hhu.edu.cn

Abstract: This paper introduces a novel ensemble adjustment Kalman filter (EAKF) that integrates a machine-learning approach. The conventional EAKF adopts linear and Gaussian assumptions, making it difficult to handle cross-component updates in strongly coupled data assimilation (SCDA). The new approach employs nonlinear variable relationships established by a deep neural network (DNN) during the analysis stage of the EAKF, which nonlinearly projects observation increments into the state variable space. It can diminish errors in estimating cross-component error covariance arising from insufficient ensemble members, therefore improving the SCDA analysis. A conceptual coupled model is employed in this paper to conduct twin experiments, validating the DNN–EAKF's capability to outperform conventional EAKF in SCDA. The results reveal that the DNN–EAKF can make SCDA superior to WCDA with a limited ensemble size. The root-mean-squared errors are reduced up to 70% while the anomaly correlation coefficients are increased up to 20% when the atmospheric observations are used to update the ocean variables directly. The other model components can also be improved through SCDA. This approach is anticipated to offer insights for future methodological integrations of machine learning and data assimilation and provide methods for SCDA applications in coupled general circulation models.

Keywords: data assimilation; machine learning; deep neural network; ensemble Kalman filter; strongly coupled data assimilation

check for updates

Citation: Wang, R.; Shen, Z. A Deep Neural Network-Ensemble Adjustment Kalman Filter and Its Application on Strongly Coupled Data Assimilation. *J. Mar. Sci. Eng.* **2024**, *12*, 108. https://doi.org/10.3390/jmse12010108

Academic Editor: Valery M. Abramov

Received: 2 December 2023
Revised: 27 December 2023
Accepted: 28 December 2023
Published: 6 January 2024

1. Introduction

As the demand for accurate weather and climate forecasting continues to rise, operational centers have recognized the importance of integrating various earth system model components, such as the atmosphere, ocean, and land, into coupled models. This integration poses challenges, particularly in the initialization of these models, where the quality of initial conditions significantly influences forecast accuracy. Coupled Data Assimilation (CDA) has emerged as a crucial method for generating initial conditions, with institutions and operational centers actively involved in advancing CDA methods [1–3]. CDA encompasses two distinct approaches: Weakly Coupled Data Assimilation (WCDA) and Strongly Coupled Data Assimilation (SCDA) [4]. In WCDA, although the background error covariance is derived from the coupled model forecast, the analysis process is carried out independently in each model component. SCDA, however, uses the full background error covariance matrix, which allows observational data from one component to influence the state variables of another component instantaneously. In theory, SCDA holds the potential to extract more information from the same observational data, maintaining a better balance between the two model components, and is the optimal CDA method [5]. However, SCDA is still in the research stage and faces a series of challenges; therefore, most operational centers are currently using WCDA.

The main challenge with SCDA is that the cross-component error covariance matrices, which are used to transfer information across components, are difficult to estimate. Han (2013) demonstrated, in a study involving a conceptual 5-variable model, that achieving superior performance with SCDA demands an exceedingly large ensemble size, typically on the order of $O(10^4)$, in contrast to WCDA [6]. Nevertheless, increasing the number of ensemble members incurs high computational costs within practical ensemble data assimilation systems. The data assimilation community has proposed various methods to enhance the effectiveness of SCDA with limited computational resources. For instance, the Leading Average Coupled Covariance (LACC) [7] method leverages the asymmetry exhibited by the ocean-atmosphere temperature correlation. It updates ocean variables by utilizing the mean of atmospheric observations and incorporates temporally leading atmospheric observations to update ocean variables, which enhances the atmospheric-ocean correlation. The covariance matrix reconditioning method [8] enhances the background error covariance matrix by modifying the original eigenvectors. The interface decomposition method [9] addresses strong coupling near the interface by artificially setting cross-component variable correlations. This approach mitigates the impact of spurious correlations and noise. Furthermore, specific methods strengthen cross-component error covariance matrices from a localized perspective, positively contributing to the assimilation process [10,11].

The above approaches have made significant progress, but they do not fully exploit the potential for the application of machine learning (ML) in data assimilation (DA). Recently, ML has found widespread applications in weather forecasting, uncertainty quantification, and data assimilation [12]. Integrating DA and ML, especially neural networks (NN), holds considerable promise to improve the accuracy and efficiency of data assimilation and model prediction. In hybrid approaches which combine DA and ML, NN can play various roles. For example, they can be employed to correct model errors through statistical correction trains using data assimilation analyses or observations [13,14]. Additionally, NN can be utilized to estimate parameters as an alternative to the augmented state approach [15,16]. Past studies have indicated that NN can serve as surrogate models by learning the data's dynamic properties. This capability allows them to replace physics-based models or the data assimilation process [17–19]. However, in these hybrid applications, ML is mostly applied to the dynamical models involved in the DA procedure rather than being directly embedded into data assimilation algorithms. Some data assimilation algorithms that incorporate a machine-learning module have been recently proposed, e.g., [20,21]. However, they did not focus on algorithms applied to SCDA.

This paper aims to exploit the capability of NN in approximating nonlinear systems and to develop a new EAKF algorithm integrated with deep neural networks (DNNs), which is particularly used in the cross-component update in SCDA. The main objectives of this paper include (a) the development of a new EAKF format in which DNN-constructed variable correlations are employed to achieve cross-component updating in SCDA; (b) the application of this DNN–EAKF approach in a conceptual model to validate its improvement for SCDA.

The organization of this paper is as follows: Section 2 introduces the method of conducting coupled assimilation using the EAKF method and subsequently presents the development of the new DNN-based EAKF algorithm through the incorporation of ML. Section 3 outlines the setup of models and twin experiments. Section 4 presents the experimental results, illustrating how the newly developed DNN–EAKF improves cross-component covariances and thus enhances strongly coupled assimilation. Finally, Section 5 provides a summary and discussion of the findings.

2. Methods

2.1. Divided State-Space Approach for CDA

The Ensemble Kalman Filter (EnKF; ref. [22]) is a widely used data assimilation method for efficiently implementing CDA. The EnKF uses an ensemble of model states to implement the update formula of the Kalman filter [23]. For a linear observation system, i.e.,

$$y = Hx,$$

H is a linear operator that maps the model state variable x into the observation space, y is the observation. The analysis scheme of the Kalman filter writes

$$x^a = x^f + PH^T(HPH^T + R)^{-1}(y^o - Hx^f), \tag{1}$$

the superscripts a, f, and o stand for analysis (posterior), forecast (prior), and observation, respectively. P denotes the background error covariance matrix, and R represents the observation error covariance matrix.

The divided state-space strategy proposed by Luo and Hoteit (2014) can be used to describe the CDA approach with EnKF [24]. For simplification, we assume that x consists of two model components $x = [x_{(a)}, x_{(o)}]$, where

$$\begin{aligned} x_{(a)} &= \{x_{(a,1)}, \ldots, x_{(a,i)}, \ldots, x_{(a,n_a)}\} \\ x_{(o)} &= \{x_{(o,1)}, \ldots, x_{(o,j)}, \ldots, x_{(o,n_o)}\} \end{aligned}$$

denote atmospheric and oceanic variables, respectively, and n_a and n_o are the number of atmospheric and oceanic variables, respectively. According to Luo and Hoteit (2014), the background error covariance matrix P in Equation (1) can also be correspondingly expressed in the form of the block matrices, i.e.,

$$P = \begin{bmatrix} P_{(aa)} & P_{(ao)} \\ P_{(oa)} & P_{(oo)} \end{bmatrix}$$

where $P_{(aa)}$ and $P_{(oo)}$ are covariances within the atmospheric and oceanic models, respectively, and $P_{(ao)}$ and $P_{(oa)}$ are cross-component error covariances.

In WCDA, the update of variables across model components is not taken into account, where the cross-component covariances are all set to zero matrices, i.e.,

$$P = \begin{bmatrix} P_{(aa)} & 0 \\ 0 & P_{(oo)} \end{bmatrix}$$

At this point, Equation (1) can be written as

$$x^a_{(a)} = x^f_{(a)} + P_{(aa)}H^T_{(a)}(H_{(a)}P_{(aa)}H^T_{(a)} + R_{(a)})^{-1}[y^o_{(a)} - H_{(a)}x^f_{(a)}] \tag{2}$$

$$x^a_{(o)} = x^f_{(o)} + P_{(oo)}H^T_{(o)}(H_{(o)}P_{(oo)}H^T_{(o)} + R_{(o)})^{-1}[y^o_{(o)} - H_{(o)}x^f_{(o)}] \tag{3}$$

where $y^o_{(a)}$ and $y^o_{(o)}$ are atmospheric and oceanic observations, respectively, and $R_{(a)}$ and $R_{(o)}$ are the corresponding observation error covariance matrices. The $H_{(a)}$ and $H_{(o)}$ are the observation operators within the corresponding models. Equations (2) and (3) indicate that the two model components carry out data assimilation independently, and the observations in each model only directly update the variables in the same model component in WCDA. The covariance matrix across the coupled components is not used.

In contrast, SCDA requires estimating the complete background error covariance matrix, which places a high demand on the number of ensemble members in the EnKF.

2.2. Ensemble Adjustment Kalman Filter with Divided State-Space

EnKF relies on ensemble statistics to compute the error covariance matrix during the data assimilation process. In practical data assimilation, various ensemble filters [22] and derivative methods (e.g., Ensemble Transformed Kalman Filter, Ensemble Square-root Kalman Filter, Ensemble Adjustment Kalman Filter [25–27] and Unscented Kalman Filter) have been proposed to implement ensemble updating in the Kalman filter. Among them, the Ensemble Adjustment Kalman Filter (EAKF) developed by Anderson [25] can decompose observations into a series of scalars and assimilate them in turn, making it well suited to coupled data assimilation problems with multiple mode components. Therefore, the present study employs the EAKF for CDA. To apply EAKF, it first assumes that vector observations can be decomposed into multiple scalars, and the scalar observations are considered independent (R is a diagonal matrix). Subsequently, it establishes iterative loops during the data assimilation process, assimilating only one scalar observation at each iteration step. The analysis serves as the a priori for the following iteration, and the process continues until all scalar observations have been assimilated.

The assimilation stage for each scalar observation comprises two steps. The initial step involves computing the observation increment based on the assumption of a Gaussian distribution. The subsequent step employs the linear regression method to regress the observation increment onto the model variables that can be incorporated into the prior states. The following provides a description of the EAKF scheme based on atmospheric observations. Further details can be found in [28].

2.2.1. Observation Increments

Initially, we denote the observation operator that projects the state vector x onto the ith atmospheric observation, represented by $y^o_{(a,i)}$, as h_i. Therefore, the projection

$$y_{(a,i)} = h_i(x). \tag{4}$$

is in the observational space. The sequential EAKF algorithm projects each member of the forecast ensemble onto the ith atmospheric observation using Equation (4), resulting in a prior ensemble of observation projections, i.e.,

$$y^f_{(a,i),k} = h_i(x^f_k), \quad k = 1, \ldots, N \tag{5}$$

Here, k in the subscript denotes the ensemble members, with a total number of N. Each ensemble member obtained through Equation (5) is a scalar value. Assuming that these members follow a Gaussian distribution, we can compute the mean $\overline{y^f_{(a,i)}}$ and the variance $(\sigma^f_{(a,i)})^2$ of the distribution from the ensemble members. Specifically,

$$\overline{y^f_{(a,i)}} = \frac{1}{N}\sum_{k=1}^{N} y^f_{(a,i),k} \tag{6}$$

$$(\sigma^f_{(a,i)})^2 = \frac{1}{N-1}\sum_{k=1}^{N}(y^f_{(a,i),k} - \overline{y^f_{(a,i)}})(y^f_{(a,i),k} - \overline{y^f_{(a,i)}})^T \tag{7}$$

Given the scalar observation $y^o_{(a,i)}$ and the observation error variance $r_{(a,i)}$ (notably, $r_{(a,i)}$ corresponds to the ith element on the diagonal of $R_{(a)}$), Bayes' rule is employed to compute the posterior probability distribution density function. This distribution conforms to the Gaussian distribution, with a variance of

$$\left(\sigma^u_{(a,i)}\right)^2 = \left[\left(\sigma^f_{(a,i)}\right)^{-2} + r^{-1}_{(a,i)}\right]^{-1} \tag{8}$$

with a mean of

$$\overline{y^u_{(a,i)}} = \left(\sigma^u_{(a,i)}\right)^2 \left[\frac{\overline{y^f_{(a,i)}}}{\left(\sigma^f_{(a,i)}\right)^2} + \frac{y^o_{(a,i)}}{r_{(a,i)}} \right] \tag{9}$$

Here, the superscript u represents the posterior value obtained from a single update. The EAKF algorithm adjusts each ensemble member to align the posterior mean and variance with the values specified by Equations (8) and (9). The posterior ensemble member in the observation space is

$$y^u_{(a,i),k} = \left(\frac{\sigma^u_{(a,i)}}{\sigma^f_{(a,i)}}\right)\left(y^f_{(a,i),k} - \overline{y^f_{(a,i)}}\right) + \overline{y^u_{(a,i)}} \tag{10}$$

Equation (10) illustrates that each ensemble member $y^u_{(a,i),k}$ is formed by shifting the mean and applying a linear contraction to the prior ensemble members. These operations of shifting and contracting ensure that the posterior sample mean equals $\overline{y^u_{(a,i)}}$, and the variance equals $\left(\sigma^u_{(a,i)}\right)^2$. For the k-th ensemble member, the observation increment is expressed as

$$\Delta y_{(a,i),k} = y^u_{(a,i),k} - y^f_{(a,i),k} \tag{11}$$

2.2.2. State-Space Increments

Given observation increments, the second step calculates the corresponding increments for each ensemble member of each state variable. For an atmospheric variable $x_{(a,j),k}$, the increment is represented as $\Delta x^{(a,i)}_{(a,j),k}$ (k indexes the ensemble member, and $j = 1, \ldots, n_a$ indexes the joint state variable throughout this study). The superscript (a,i) indicates that the increment is associated with the observation $y^o_{(a,i)}$.

The serial EAKF algorithm requires assumptions about the prior relationship among joint state variables, encompassing both observed and unobserved variables. This algorithm assumes that the prior distribution follows a Gaussian distribution. This assumption is equivalent to assuming that a local least-squares fit to the prior ensemble members captures the relationship among the joint state variables.

Figure 1a replicates the straightforward illustration from Anderson (2003) [28], depicting the relationship between update increments for a state variable x and an observation variable y. The observation variable is linked to the state variable through a typically nonlinear operator g. As observation increments have been determined using Equation (11), the corresponding increments for the state variable can be calculated through a global least-squares fit. Thus, the increment from the observation $y^o_{(a,i)}$ is given by

$$\Delta x^{(a,i)}_{(a,j),k} = \frac{\sigma_{x_{(a,j)},y}}{(\sigma^f_{(a,i)})^2} \Delta y_{(a,i),k}, \qquad j = 1, 2, \ldots, n_a \tag{12}$$

Here, $\sigma_{x_{(a,j)},y}$ signifies the covariance between $x_{(a,j)}$ and $y^f_{(a,i)}$, calculated from ensemble members, while $(\sigma^f_{(a,i)})^2$ represents the prior ensemble variance computed using Equation (7).

Adding $\Delta x^{(a,i)}_{(a,j),k}$ to $x^f_{(a,j),k}$ results in the updated analysis field $x^u_{(a,j),k}$. Subsequently, iterate over j to update all atmospheric variables using the same atmospheric observation.

$$x^u_{(a,j),k} = x^f_{(a,j),k} + \Delta x^{(a,i)}_{(a,j),k}, \quad j = 1, 2, \ldots, n_a \tag{13}$$

It is important to note that if the localization method is employed, the term $\Delta x^{(a,i)}_{(a,j),k}$ in Equation (12) should be multiplied by the localization factor ρ, which is linked to the distance between the locations of $x^f_{(a,j)}$ and $y^o_{(a,i)}$. For simplicity in the discussion, we refrain from utilizing the localization method in the experiments.

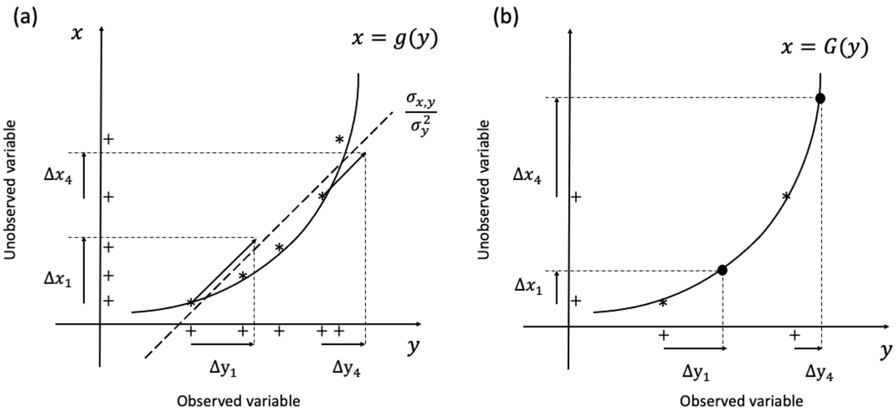

Figure 1. Schematic diagram of the state variable update algorithm in EAKF (**a**) and DNN-EAKF (**b**), where g or G is the nonlinear relationship between the observed variable y and the unobserved variable x, with "*" representing ensemble members and "+" representing the projection of ensemble members on the x/y axis.

For an oceanic variable $x_{(o,j),k}$ that requires updating through cross-component correlation using atmospheric observations, the same linear regression approach can still be employed to compute the increment for each ensemble member, as follows:

$$\Delta x^{(a,i)}_{(o,j),k} = \frac{\sigma_{x_{(o,j)},y}}{(\sigma^f_{(a,i)})^2} \Delta y_{(a,i),k}, \quad j = 1, 2, \ldots, n_a \tag{14}$$

Ocean variables can be updated using the same process, where

$$x^u_{(o,j),k} = x^f_{(o,j),k} + \Delta x^{(a,i)}_{(o,j),k}, \quad j = 1, 2, \ldots, n_o \tag{15}$$

Here, n_o denotes the number of ocean variables.

Nevertheless, certain studies have indicated that the methods outlined in Equation (14) and (15) require very large ensembles for accurately estimating cross-component correlation coefficients. Otherwise, the increments derived from Equation (14) might be significantly biased, resulting in erroneous assimilation effects in Equation (15) [6]. This is primarily due to the strong nonlinear correlations among variables from different components within the coupled model, making it challenging for regression methods based on the assumption of local linearity to precisely estimate their correlation coefficients (Figure 1a) and necessitating a considerable number of members to achieve the desired effect.

2.2.3. DNN-Based State-Space Increments for EAKF

In this study, we introduce DNN to model relationships between cross-component variables. The DNN is an artificial neural network characterized by multiple hidden layers designed for learning and representing complex nonlinear relationships. The primary strength of DNN lies in its efficient ability to capture and model complex, nonlinear relationships in data.

As an example, using ocean and atmosphere variables, we introduce a projection operator Π from the atmosphere to the ocean, expressed as $x_{(o)} = \Pi(x_{(a)})$. This projection operator facilitates the representation of cross-component inter-variable relationships through an NN model trained on data derived from background integration. The nonlinear relationship based on neural networks is expressed as $\mathcal{G}(\Pi(x_{(a)}),\theta)$, where the vector θ represents the trainable parameters of the neural network. The optimal weights are determined through an iterative process of minimizing the loss function. $x_{(o)}$ serves as the label for the training set. The function $\mathcal{G}(\Pi(x_{(a)}),\theta)$ can be solved using the following optimization problem:

$$L(\theta) = \sum_{i=1}^{N_f} \left\| \mathcal{G}^{(i)}\left(\Pi\left(x_{(a)}\right),\theta\right) - x_{(o)} \right\|^2_{P_k^{-1}} \tag{16}$$

Here, N_f denotes the length of the training set, representing the minimization of the error between predicted and true values. P_k is a symmetric semi-positive definite matrix defining the paradigm $\|x\|^2_{P_k^{-1}} = x^T P_k^{-1} x$, equivalent to the error covariance matrix of the NN model.

By employing integration or reanalysis data, we can train the model parameters to derive the nonlinear function $\mathcal{G}(\Pi(x_{(a)}),\theta)$, utilizing atmospheric variables to predict oceanic variables. This function is subsequently employed to project a priori and a posteriori values from the atmospheric component to the oceanic component, facilitating the computation of variable increments within the oceanic model. This can be expressed using

$$\Delta x_{(oj),k} = \tilde{G}(x^u_{(a),k}) - \tilde{G}(x^f_{(a),k}), j = 1,\ldots,n_o \tag{17}$$

where $\tilde{G}$, as the function of $x_{(a)}$, denotes the nonlinear function $\Pi(x_{(a)})$ with the optimal parameter derived by solving Equation (16).

Figure 1b depicts a schematic of the algorithm. In this context, updates of the unobserved variables are obtained not through linear regression but by employing a nonlinear model trained by NN. Finally, utilizing Equation (15), it is possible to obtain the a posteriori values of the oceanic component. Again, localization methods can also be applied in this stage.

Due to the reliance on DNN to establish nonlinear relationships between variables, we term the newly proposed method Deep Neural Network-Ensemble Adjustment Kalman Filter (DNN–EAKF). Figure 2 presents a flow chart of the conventional EAKF method and the DNN–EAKF developed in this section. Here, we take atmospheric observations and ocean and atmospheric model variables as examples, but they can also be applied to more general situations. It is also noteworthy that DNN–EAKF is only applied for cross-component updates, while intra-component updates still use conventional EAKF.

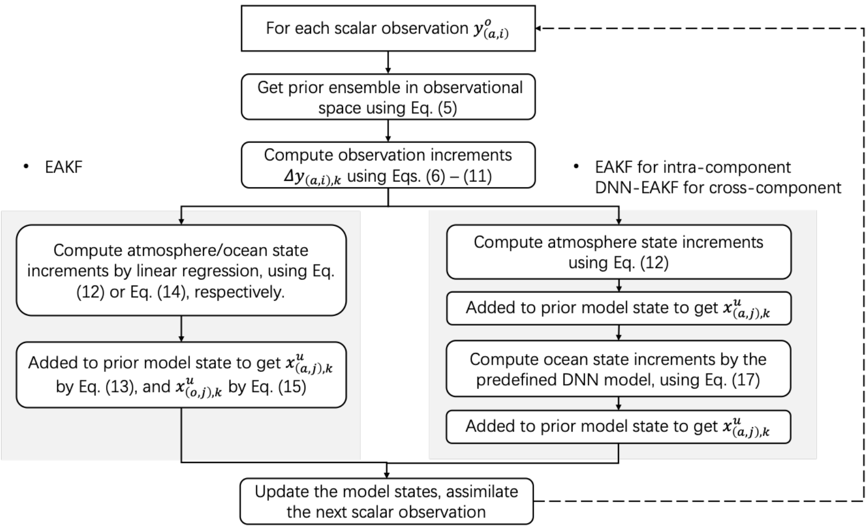
For each scalar observation $y^o_{(a,i)}$
Get prior ensemble in observational space using Eq. (5)
Compute observation increments $\Delta y_{(a,i),k}$ using Eqs. (6) – (11)
• EAKF
• EAKF for intra-component
DNN-EAKF for cross-component
Compute atmosphere/ocean state increments by linear regression, using Eq. (12) or Eq. (14), respectively.
Added to prior model state to get $x^u_{(a,j),k}$ by Eq. (13), and $x^u_{(o,j),k}$ by Eq. (15)
Compute atmosphere state increments using Eq. (12)
Added to prior model state to get $x^u_{(a,j),k}$
Compute ocean state increments by the predefined DNN model, using Eq. (17)
Added to prior model state to get $x^u_{(a,j),k}$
Update the model states, assimilate the next scalar observation

Figure 2. Flow chart of the conventional EAKF method (**left** route) and the DNN–EAKF developed in this section (**right** route).

3. Model and Experimental Settings

3.1. Numerical Model

The numerical model employed in this study is a conceptual coupled model extensively utilized in prior research to evaluate the efficacy of data assimilation methods (e.g., [6,29–32]). This coupled model comprises a fast atmosphere, a slow upper ocean, and a significantly slower deep ocean with an idealized sea ice component. Although the simple coupled model may lack the physical complexity of the coupled circulation model, it effectively characterizes interactions among multiple time-scale components in the climate system [33] and adeptly captures certain challenges in SCDA.

The equation for this low-order coupled model is

$$\begin{aligned}
&\dot{x}_1 = -\sigma x_1 + \sigma x_2 \\
&\dot{x}_2 = -x_1 x_3 + (1 + c_1\omega)\kappa x_1 - x_2 \\
&\dot{x}_3 = x_1 x_2 - b x_3 \\
&O_m \dot{\omega} = c_2 x_2 + c_3 \eta + c_5 \omega\eta - O_d \omega + S_m + S(t) - c_7 \varphi_{t-1} \\
&\Gamma \dot{\eta} = c_4 \omega + c_6 \omega \eta - O_d \eta \\
&\varphi_t = \Phi(x_2, \omega, \varphi_{t-1}),
\end{aligned}$$

where the six model variables represent the atmosphere, the ocean, and the sea ice x_1, x_2, and x_3 are for the atmosphere (hereafter denoted by $x_{1,2,3}$ if present together), ω is for the slab ocean, η is for the deep-ocean pycnocline, and φ is for the sea ice concentration. The dots above the variables indicate time trends (time derivatives). In this simple system, the seasonal period is defined as 10 nondimensional model time units (TUs, 1 TU = 100 time steps, given $\Delta t = 0.01$), and a model year (decade) is 10 (100) TUs. The atmosphere model is Lorenz's chaotic model [34], the standard values of the original parameters σ, κ, and b are, respectively, 9.95, 28, and 8/3, and the atmospheric time scale is defined as 1 TU. The coupling between the fast atmospheric and the slow ocean is achieved by choosing the values of the coupling coefficients c_1 and c_2, which denote the ocean-to-atmosphere and the atmosphere-to-ocean forcing, respectively. The parameters c_3 and c_5 denote the linear forcing of the deep ocean and the nonlinear interaction of the upper ocean with the deep ocean. O_m is the ocean heat capacity, while O_d denotes the damping coefficient of the flat ocean variable ω. Their values define that the time scale of the ocean variable ω is much slower than the atmosphere, e.g., $(O_m, O_d) = (10, 1)$ defines the oceanic time scale to be approximately 10 times that of the atmosphere. In addition, the model uses the term $S(t) = S_m + S_s \cos\left(2\pi t / S_{\text{pd}}\right)$ to simulate constant and seasonal forcing of the "climate" system. The parameter c_7 denotes the coupling coefficient between sea ice and the slab ocean. In the pycnocline model, η represents the anomaly of the ocean pycnocline depth, with its trend equation derived from a binomial equilibrium model of the latitudinal time-averaged specific gravity pycnocline, interacting with ω. The constant of proportionality is denoted as Γ, while c_4 and c_6 represent the linear forcing of the upper ocean and the nonlinear interaction of the upper ocean with the deep ocean. Finally, the sea ice model takes the form of a straightforward nonlinear function that maps enthalpy space to the sea ice concentration space. In this context, "sea ice" φ influences the atmosphere solely through the interaction of the ocean variable ω and the atmospheric variable x_2.

To solve the assimilation problem caused by the discontinuity in the distribution of sea ice concentration, Zhang et al. (2013) introduced a nonlinear function of enthalpy ($H = c_8 x_2^2 + c_9(\omega - 10)^2 + c_{10}\varphi_{t-1}$) to define the sea ice medium [33], in which the nonlinear transformation function from enthalpy to ice concentration is

$$\varphi = \Phi(H) = \begin{cases} 0, & H > H_{\text{ig}} \\ 1, & H < H_{\text{im}} \\ 0.5\left[e^{-(H - H_{\text{im}})^{-1}} + e^{-\left(H_{\text{ig}} - H\right)/H_0}\right], & H_{\text{im}} \le H \le H_{\text{ig}} \end{cases},$$

The H_{ig} and H_{im} represent the thresholds for the ice generation and maintenance points, while H_0 is used to adjust the shape of the curve, distributed between 0 and 1. It also has both x_2 and ω time scales according to the formulation.

Referring to Han et al. (2013) [6], the parameter values of the true-value model are (σ,κ, b, c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8, c_9, c_{10}, O_m, O_d, S_m, S_s, S_{pd}, Γ, H_{ig}, H_{im}, H_0) = (9.95, 28, 8/3, 0.1, 1, 0.01, 1, 0.01, 0.01, 0.01, 0.1, 0.1, 0.1, 10, 10, 10, 10, 1, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 100, 50, 10, 80). We integrated the model using a fourth-order Runge–Kutta scheme, starting with the initial conditions (x_1 , x_2 , x_3 , ω , η , φ) = (0 , 1 , 0 , 0 , 0 , 0 , 0), and using the values after spin-up over 2500 TUs as the true initial values. Figure 3 shows the time series of the three atmospheric and two oceanic variables, as well as the sea ice variable, and it can be observed that the three atmospheric variables have attractor characteristics. The x-axis of Figure 3b,c uses a different time scale, revealing that the variability of the oceanic variables is about 1/10 of that of the atmospheric variables.

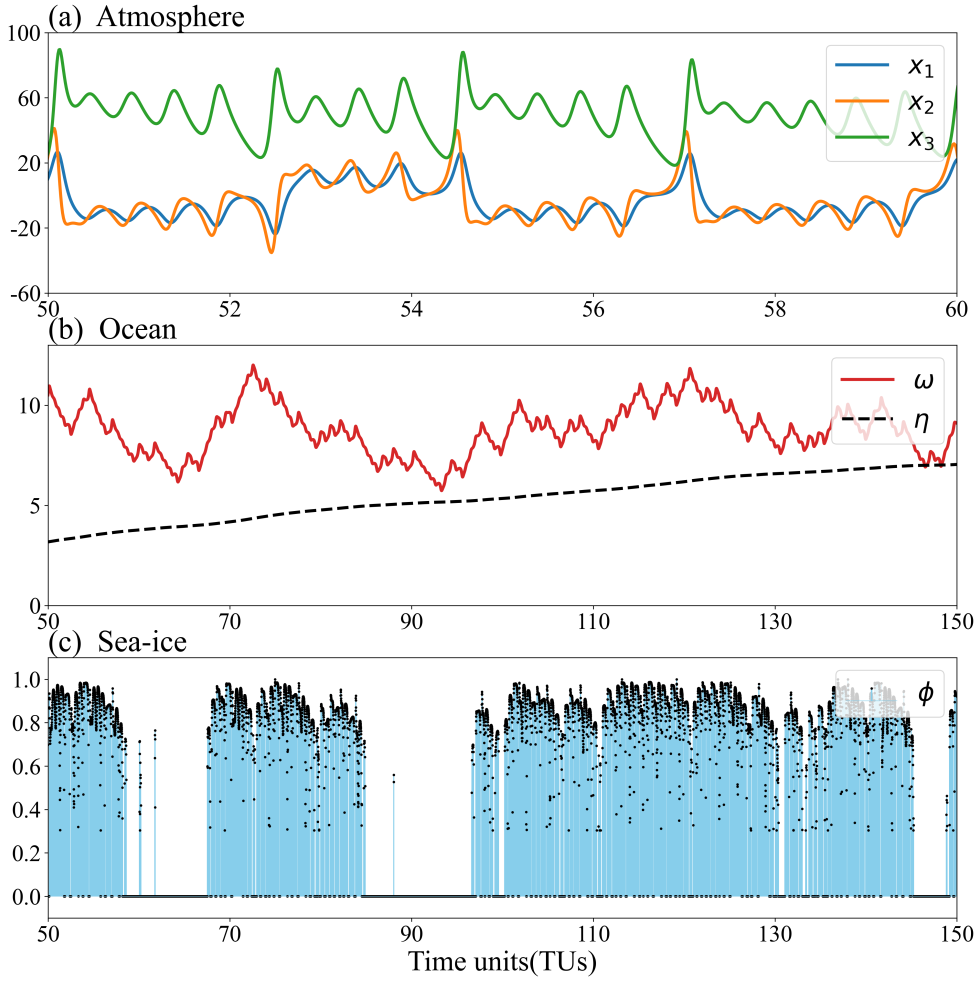

Figure 3. The model state values of the atmosphere (**a**) $x_{1,2,3}$; ocean (**b**) ω and η; and sea ice (**c**) φ, with 0, 1, 0, 0, 0 and 0 as the initial conditions for x_1, x_2, x_3, ω, η and φ. We showed the time series of the 3 components during the period of 50–60 TUs, 50–150 TUs and 50−150 TUs, respectively.

3.2. Neural Network Model

We utilize model integration data to train an NN model aimed at establishing nonlinear relationships between atmospheric and oceanic variables. Specifically, for the coupled model employed in this study, we formulate the nonlinear relationship from the atmospheric variable $\boldsymbol{X} = [x_1, x_2, x_3]$ to each of the oceanic and sea ice variables: ω, η, and φ.

Taking ω as an example, the objective of ML training is to construct a neural network $\tilde{G}_W$ to predict ω using the atmospheric variable $x_{(a)}$, with W representing its weight. The optimal weights are determined through the minimization of the loss function during the training phase. To acquire the training data, we conduct a background integration of 5000 TUs for the model starting from a random initial value, with an integral step size of $\Delta t = 0.01$. These data are divided into 5000 TUs, comprising input–output pairs of atmospheric variables ($x_{1,2,3}$) and oceanic variables (ω) at corresponding moments. Among these, 4000 TUs are allocated for model training, 800 TUs for validation and hyperparameter tuning during the training period, and the remaining 200 TUs are dedicated to evaluating the model's robustness without any overlap among the three sequences.

Based on different assumptions, we constructed three DNN models as outlined below:

In the first model, we utilize atmospheric variables to predict concurrent oceanic variables, naming this model the Single-Instant Predictor (SIP). The training involves a three-layer fully connected network model; refer to Figure 4a for the schematic neural network structure. The training objective is to achieve the desired nonlinear model, expressed as $\omega(t) = \tilde{G}(\boldsymbol{X}(t))$, where t indicates the time step.

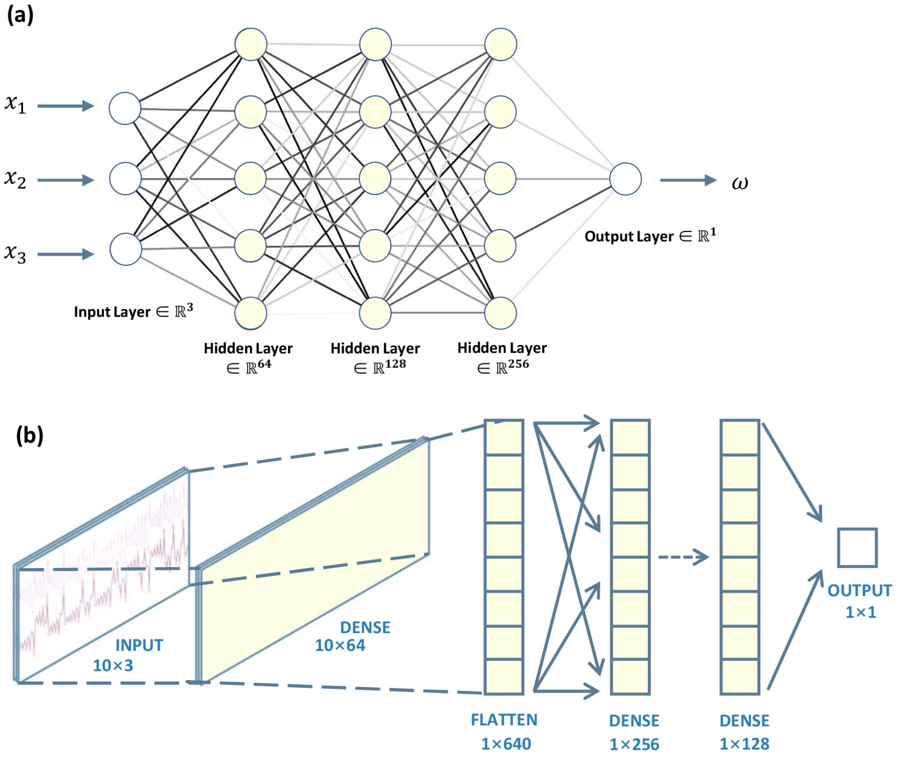
(a)
x_1
x_2
x_3
ω
Input Layer $\in \mathbb{R}^3$
Hidden Layer $\in \mathbb{R}^{64}$
Hidden Layer $\in \mathbb{R}^{128}$
Hidden Layer $\in \mathbb{R}^{256}$
Output Layer $\in \mathbb{R}^1$
(b)
INPUT 10×3
DENSE 10×64
FLATTEN 1×640
DENSE 1×256
DENSE 1×128
OUTPUT 1×1

Figure 4. Schematic diagram of the neural networks. (**a**) A three-layer fully connected neural network for two single-step prediction models with input and output sizes of 3 and 1, respectively; (**b**) A fully connected neural network for a multi-step prediction model with input and output sizes of 10 × 3 and 1, respectively.

In accordance with findings by Lu et al. (2015), indicating that utilizing atmospheric observations with lead times can significantly enhance analysis quality in WCDA compared to SCDA using a small ensemble size [7], our second model associates the oceanic variables with the previous atmospheric variables. Specifically, we construct the model to predict the current oceanic variable using the atmospheric variable 0.2 TUs ahead of time. We term this model a Single-Step Leading Predictor (SLP). The network is the same as SIP, also refer to Figure 4a, and the training results in a target nonlinear model expressed as $\omega(t) = \widetilde{G}(\boldsymbol{X}(t-0.2))$.

Building on Lu et al. (2015) strategy of averaging time-leading atmospheric variables to construct ocean-atmosphere covariance, which reduces noise arising from disparate variability in the atmosphere and ocean, our third model utilizes all atmospheric variables from the 10 consecutive steps to predict ocean variables in the final step. Termed the

Multi-Leading Predictor (MLP), the network model is depicted in Figure 4b, trained to achieve the target nonlinear model $\omega(t) = \widetilde{G}(\boldsymbol{X}(t-9), \ldots, \boldsymbol{X}(t))$. Importantly, the inputs in each data pair include atmospheric variables from the 10-step model integration, while the outputs represent oceanic variables from the final step. Consequently, the volume of data for both training and validation is only 1/10th of that used for the first two models.

We use the three aforementioned models to establish the relationship between $\boldsymbol{X}$ and ω. The models' parameters are optimized using the Adam algorithm, and the loss function is the Huber loss over the training dataset, which is made of background snapshots. The training consists of 200 epochs with an adaptive learning rate (initial learning rate sets 1×10^{-3}) and batch size of 50. After the entire training step, we keep the model that yields the lowest loss over the validation dataset. The three aforementioned models to establish the relationship between $\boldsymbol{X}$ and ω, the stabilized prediction results are shown in Figure 5. Figure 5a–c illustrate the performances of the three models on the same test set, where the red line represents the true value, while the blue, green, and violet lines correspond to the predicted ω-values by the three models, respectively. It can be observed that all three models can roughly simulate the trend of the true value, and for the SIP, some extreme values appear due to the large variability difference between the atmospheric and oceanic variables; the stability of the SLP (b) is significantly improved compared to (a); and the MLP (c) achieves the best prediction result. This also implies that the use of training data related to the leading-averaged atmospheric variables enhances the accuracy of sea-air predictions.

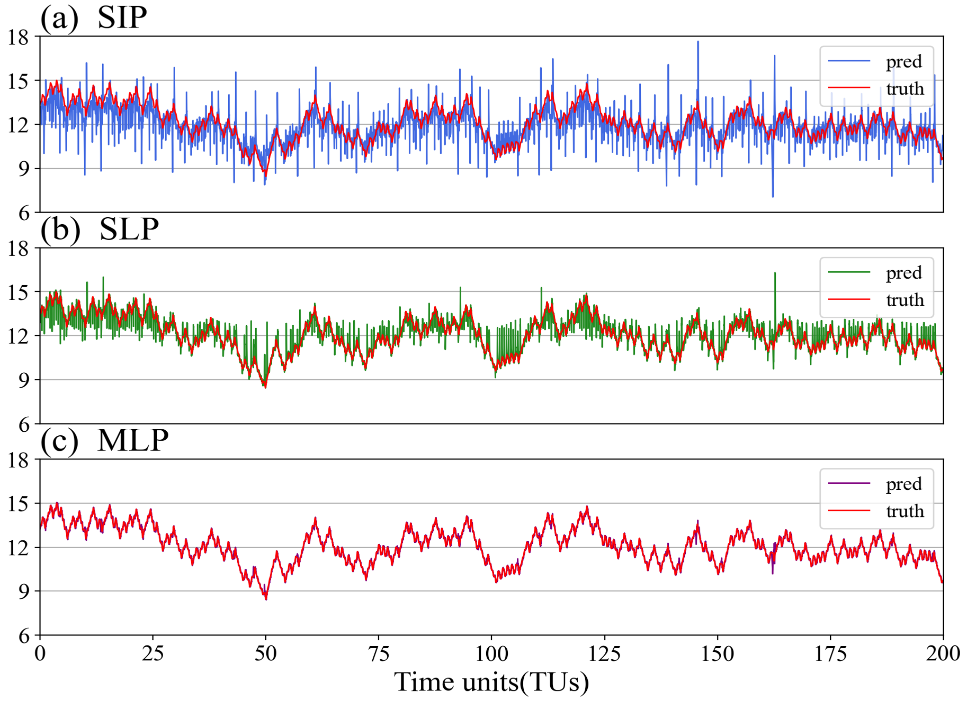

Figure 5. Prediction effects of Single-instant Predictor (**a**), Single-leading Predictor (**b**) and Multi-leading Predictor (**c**) on the test set .

We have employed a similar approach to train the relationship between $\boldsymbol{X}$ and the deep-ocean pycnocline variable η, as well as the sea ice variable φ with MLP, as shown in Figure 6. The results show that the relationship between the fast-varying atmospheric variables and the slow-varying pycnocline variables is notably weak, rendering the prediction of η with the atmosphere nearly impossible. On the other hand, the prediction of sea ice with atmospheric variables proves to be difficult. This discrepancy can be attributed to the relatively weak connection between these variables in the model equations. We, therefore, focus on the SCDA of atmospheric observations to the sea-surface variable ω

in the assimilation experiments below, which is also consistent with the idea of interface decomposition proposed by [9].

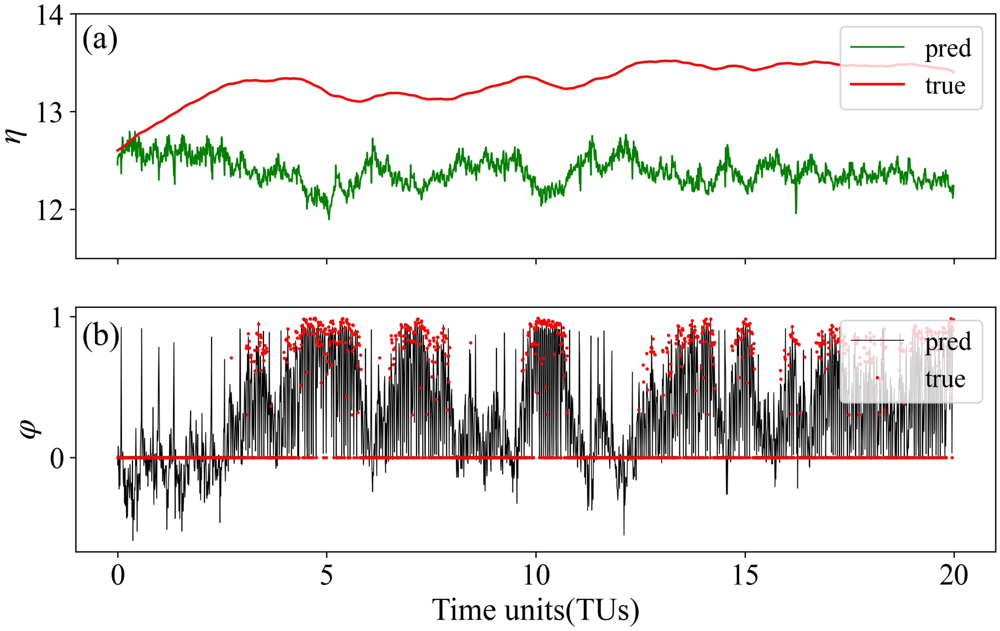

Figure 6. MLP model for η (**a**) and φ (**b**).

3.3. Data Assimilation Experiment Settings

The true values and observations used as reference are generated by integrating the coupled model with standard parameter values (cf. Section 3.1). The integration begins with the true initial values obtained in Section 3.1, and the model integration step is $\Delta t = 0.01$ TUs, spanning a total of 100 TUs for the entire experiment. Observational data are generated by adding random noise, following a specific distribution, to the true values. To simulate real-world conditions, we assume that atmospheric, sea-surface, and sea ice variables can be observed at specific time intervals, whereas the pycnocline variable η is unobservable. Following the setup of Zhang [31], we assume that observation errors for the atmospheric variables $x_{1,2,3}$ all follow a Gaussian distribution with a standard deviation of 2. The observation errors for ω and φ are assumed to follow Gaussian distributions with standard deviations of 0.5 and 0.1, respectively. Additionally, to simulate mode errors, we introduce biased coupled modes in both the background integration and assimilation experiments. Here, all physical parameters are perturbed from the reference parameters with a 1% random error.

Assimilation experiments were conducted to compare the performance of conventional EAKF and DNN–EAKF in SCDA. Prior studies indicate that high-frequency atmospheric observations positively impact oceanic variables, while low-frequency oceanic observations struggle to adjust atmospheric variables [6,35]. Hence, our emphasis is on evaluating SCDA concerning atmospheric observations while using WCDA to assimilate ocean and sea ice observations. Three CDA frameworks can be established for atmospheric observations: WCDA, SCDA at the interface (SCDA-I), and fully SCDA (SCDA-F). The influence of observations on various variables is illustrated in Figure 7. Among them, SCDA-I utilizes atmospheric observations to update the sea-surface variable ω, which references [9].

As shown in Figure 5, a DNN can efficiently establish a nonlinear relationship between atmospheric variables $x_{1,2,3}$ and the sea-surface variable ω. However, establishing a relationship between atmospheric variables and pycnocline or sea ice variables remains challenging (Figure 6). Therefore, we employ DNN–EAKF to update ω using atmospheric observations in SCDA-I (black box in Figure 7b). We then compare the results with those of the three CDA experiments using the conventional EAKF method.

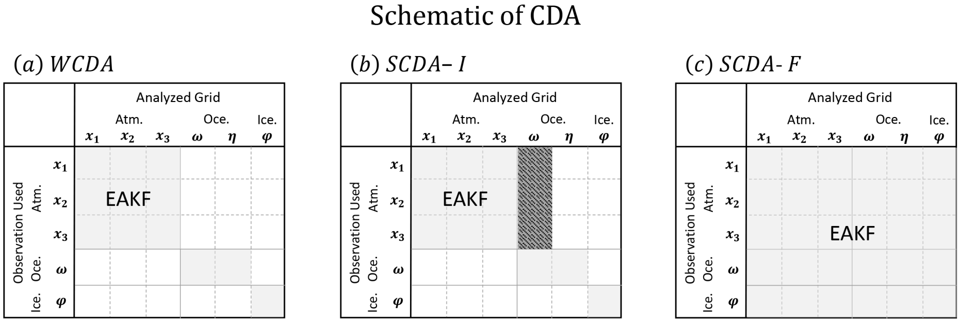

Figure 7. Three frameworks for coupling assimilation: WCDA (**a**), SCDA at the interface (**b**) and fully SCDA (**c**). The horizontal axis represents the observed variables, and the vertical axis represents the variables affected by each observation. The dark gray shadows in SCDA-I (**b**) represent the DNN–EAKF method obtained using conventional EAKF or DNN-based training models in the atmospheric observation effect ω variable.

In the subsequent discussion, we initially present the results of assimilating only atmospheric observations and subsequently extend our analysis to encompass the assimilation of all available observations.

4. Results

4.1. Atmosphere Observations

In the initial scenario, the focus is on assimilating exclusively atmospheric observations into the coupled model, employing various CDA frameworks and methodologies. Various assimilation intervals (e.g., assimilating atmospheric observations every 0.1 or 0.2 TUs) and multiple ensemble sizes (with N representing the ensemble member size, ranging from 10 to 50) are explored. The assessment of data assimilation results includes comparing true values using metrics such as root-mean-squared error (RMSE) and anomaly correlation coefficient (ACC). In this context, RMSE and ACC are defined as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{K}\sum_{i=1}^{K}\left(\overline{x_i^a} - x_i^{\mathrm{true}}\right)^2} \tag{18}$$

and

$$\mathrm{ACC} = \frac{\frac{1}{K-1}\sum_{i=1}^{K}\left(\overline{x_i^a} - \overline{x^a}\right)\left(\overline{x_i^{true}} - \overline{x^{true}}\right)}{\sqrt{\frac{1}{K-1}\sum_{i=1}^{K}\left(\overline{x_i^a} - \overline{x^a}\right)^2}\sqrt{\frac{1}{K-1}\sum_{i=1}^{K}\left(\overline{x_i^{true}} - \overline{x^{true}}\right)^2}} \tag{19}$$

In these equations, K denotes the time steps for a state variable x, while $\overline{x^a}$ and x^{true} signify the ensemble mean of the analysis and true values of variable x, respectively. To ensure the reliability of the conclusions, we utilized the results from the last 30 TUs for calculating RMSE and ACC.

To mitigate the impact of randomness in the outcomes, each experiment was replicated 10 times with different initial perturbation values. The final results were determined based on the mean values of RMSE and ACC obtained from these ten experiments.

In this scenario, we examine the outcomes related to the atmospheric variable x_2 and the oceanic variable ω. Figure 8 depicts histograms illustrating the RMSE distributions for each method. The bar values denote the mean of 10 replicate experiments, and the error bars indicate their standard deviation.

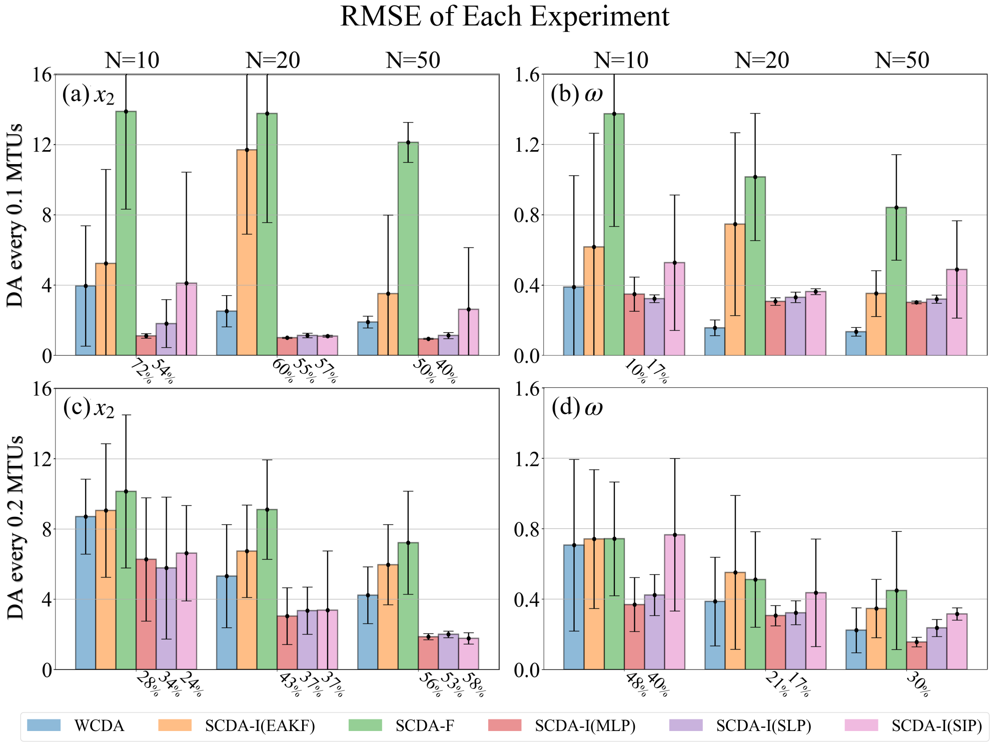

Figure 8. The RMSEs of WCDA (blue), SCDA-I(EAKF) (orange), SCDA-F (green), SCDA-I(MLP) (red), SCDA-I(SLP) (violet), and SCDA-I(SIP) (pink) for the atmospheric variable x_2 when the atmospheric observation interval is 0.1 TUs (**a**) and 0.2 TUs (**c**), within the [70, 100] TUs timeframe; (**b**,**d**) same as (**a**,**c**), but for the ocean variable ω. The error bars represent the standard deviation of the RMSE of 10 replicate experiments for each coupled method; the ratios beneath the bar of DNN-EAKF are the error reduction rates compared to WCDA (blue).

Comparing RMSE, it is evident that, in the realm of CDA utilizing EAKF, SCDA-F (green) exhibits poor performance, while WCDA (blue) demonstrates superior performance. Additionally, their assimilation effectiveness improves with larger ensemble sizes and more frequent observations. This indicates that introducing cross-component error covariance through linear approximation may degrade state estimation when there are insufficient ensemble members, consistent with findings in [6]. The detrimental impact of an increased frequency of atmospheric observations on the assimilation performance of SCDA-F (green) is conspicuous. This implies that poorly estimated cross-component error covariances can accumulate adverse effects when rapidly incorporating atmospheric observation information.

Regarding the ocean variable ω in Figure 8b,d, DNN–EAKF consistently outperforms the conventional EAKF approach in SCDA-I (orange). It indicates that, even with smaller ensembles, atmospheric observations can accurately adjust ocean variables through nonlinear mapping, thanks to the enhanced signal-to-noise ratio of the cross-component error covariance. In line with the diverse behaviors of different models illustrated in Figure 5, SCDA-I(MLP) (red), exhibiting the highest prediction accuracy, performs optimally in most cases. However, the large standard deviation of the results from the 10 replicate experiments reveals that the predictive effect of SCDA-I(SIP) (pink) is not sufficiently stable, limiting the method’s performance.

Additionally, it is observed that SCDA-I using DNN–EAKF surpasses the assimilation effect of WCDA in some instances, particularly with an observation interval of 0.2 TUs. This implies a positive updating effect of atmospheric observations on oceanic variables through DNN–EAKF. We define the relative error reduction rate r as

$$r = \frac{\overline{RMSE}_{WCDA} - \overline{RMSE}_{DNN\tilde{}EAKF}}{\overline{RMSE}_{WCDA}},$$

denoting the relative error reduction of SCDA-I using DNN–EAKF compared to WCDA and representing the improvement effect from the cross-component update based on DNN–EAKF. The value of r is indicated in Figure 8 beneath the bar where SCDA-I using DNNs outperforms WCDA. It highlights that the error reduction of DNN–EAKF over WCDA becomes more prominent with an extended atmospheric observation interval and a smaller ensemble size. This signifies situations where DNN–EAKF holds a more significant advantage, namely when the problem is more nonlinear and the ensemble size is limited.

Figure 8c shows a notable increase in the relative error reduction rate for the atmospheric variables with an expanding ensemble when the observation interval is 0.2 TUs. This is attributed to the substantial improvement that DNN–EAKF brings to the oceanic variables in this scenario, resulting in a decrease in oceanic error with the increasing ensemble size. It can be inferred that the coupling with the improved ω contributes to improving the accuracy of x_2 in SCDA-I using DNN–EAKF.

Figure 9 displays the ACC corresponding to the results shown in Figure 8, reaffirming the same conclusion. It should be noted that the ratio labeled is ACC growth instead of RMSE reduction. Clearly, SCDA-I based on DNN–EAKF, especially when utilizing the MLP model, consistently produces significantly enhanced assimilation results compared to WCDA. The advantage is more pronounced in scenarios characterized by stronger nonlinearity and smaller ensemble sizes.

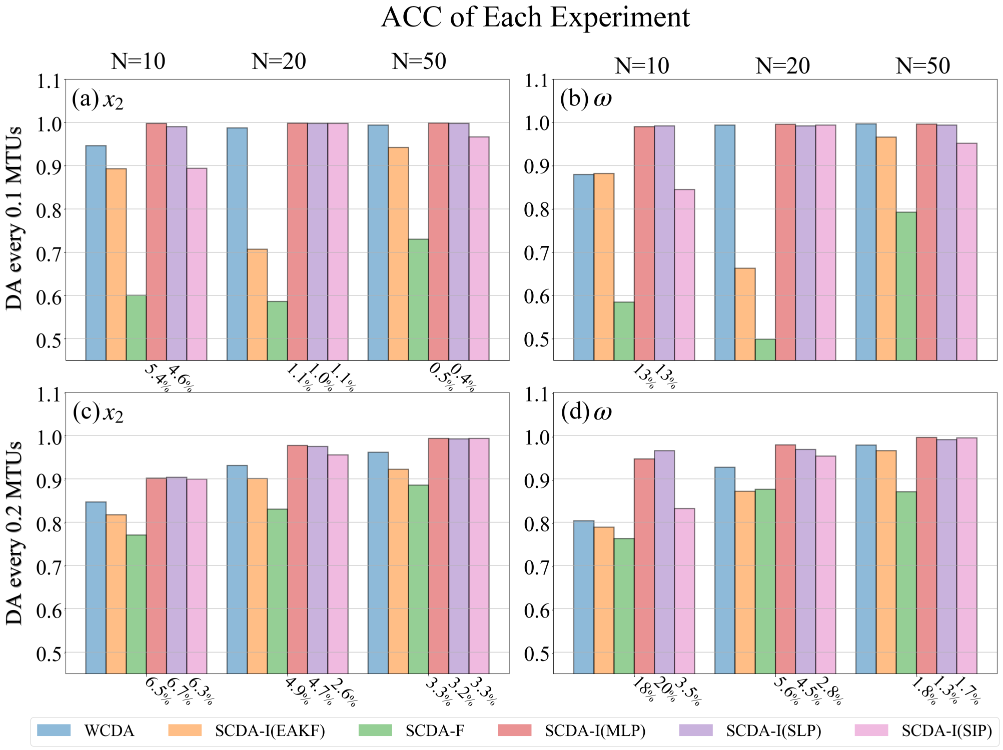

Figure 9. Same ACCs of WCDA (blue), SCDA-I(EAKF) (orange), SCDA-F (green), SCDA-I(MLP) (red), SCDA-I(SLP) (violet), and SCDA-I(SIP) (pink) for the atmospheric variable x_2 when the atmospheric observation interval is 0.1 TUs (**a**) and 0.2 TUs (**c**), within the [70, 100] TUs timeframe; (**b**,**d**) same as (**a**,**c**), but for the ocean variable ω. The error bars represent the standard deviation of the RMSE of 10 replicate experiments for each coupled method; the ratios beneath the bar of DNN-EAKF are the ACC value growth rates compared to WCDA (blue).

It is intriguing to further investigate how DNN–EAKF addresses nonlinearities. Figure 10 illustrates the probability distributions of x_2 and ω (i.e., climatological state distributions) in the mean of analysis field using 10 ensemble members at an observation interval of 0.2 TUs. Once again, we rely on the results from the last 30 TUs and compare the climatological distributions of WCDA, SCDA-I (MLP), and the true values. Notably, the climatological distributions of the atmospheric variables do not differ significantly between the three methods. However, for the oceanic variables, the climatological distribution of EAKF results differs significantly from the true climatological distribution, but the DNN–EAKF aligns with it more closely. It suggests that the method excels in handling nonlinear and non-Gaussian problems, shedding light on the underlying reasons for its advantages.

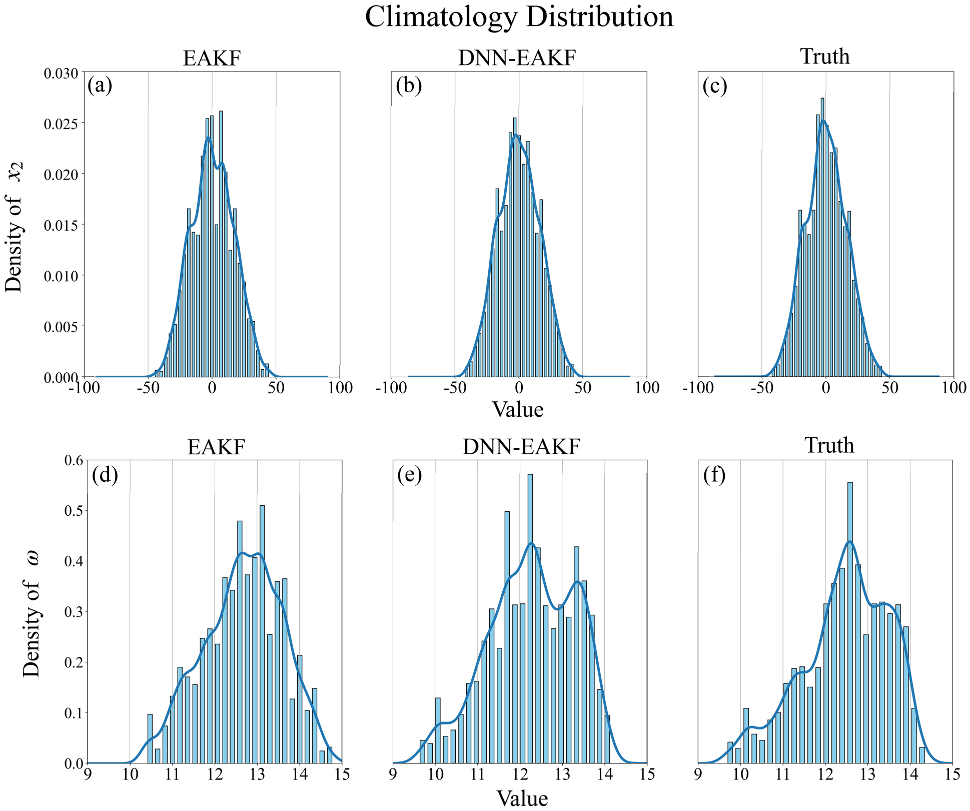

Figure 10. The probability distributions of the atmospheric variable x_2 for the EAKF (WCDA) (**a**), DNN−EAKF (SCDA−I(MLP)) (**b**), and true values (**c**) during the period [70 TUs, 100 TUs] are shown, respectively; (**d**–**f**) same as (**a**–**c**), but for the ocean variable ω. This analysis is based on an experiment with 10 ensemble members assimilating atmospheric observations every 0.2 TUs.

4.2. Multiple Observations

Experiments in Section 4.1, exclusively assimilating atmospheric observations, demonstrated that SCDA-I using DNN–EAKF effectively improves cross-component updates. It results in a more stable and accurate model state compared to EAKF, particularly in conditions with pronounced nonlinearities, such as those with low assimilation frequencies and few ensemble members. In this section, we investigate the broader impact of DNN–EAKF on the overall variables of the coupled model, considering multiple observations, including atmosphere, sea surface, and sea ice. A more realistic scenario is considered, where oceanic observations are less frequent than those for the atmosphere. Specifically, we assume an observation interval of 0.1 TUs for the atmosphere and 0.5 TUs for the oceans and sea ice. This results in atmospheric observations directly updating ocean variables more frequently than oceanic observations in SCDA-F and SCDA-I. However, inaccurate cross-component error covariances can significantly degrade assimilation results compared to WCDA.

We calculate the RMSE at each step using the experimental results of the 10 ensemble members compared to the true values and present them in Figure 11, focusing on the impact of SCDA-I (MLP). To mitigate randomness, we computed the average RMSE over 10 repeated experiments. For the presentation, we applied a smoothing process to the time series of RMSE, using a moving average with a window size of 1 TU (or 100 steps). Consistent with findings from experiments assimilating only atmospheric observations, SCDA-I (MLP) effectively assimilates both x_2 and ω, surpassing the performance of the three coupled assimilation frameworks using the conventional EAKF. Although SCDA-F's influence on the assimilation of these two variables is weaker than that of WCDA and SCDA-I, it remains within an acceptable range due to the presence of ocean observations constraining the ocean variables. The enhancements in ocean variable assimilation achieved by SCDA-I are also manifested in η and φ (Figure 11 shows the corresponding enthalpies H), indicating improved assimilation results for the deep-sea and sea ice.

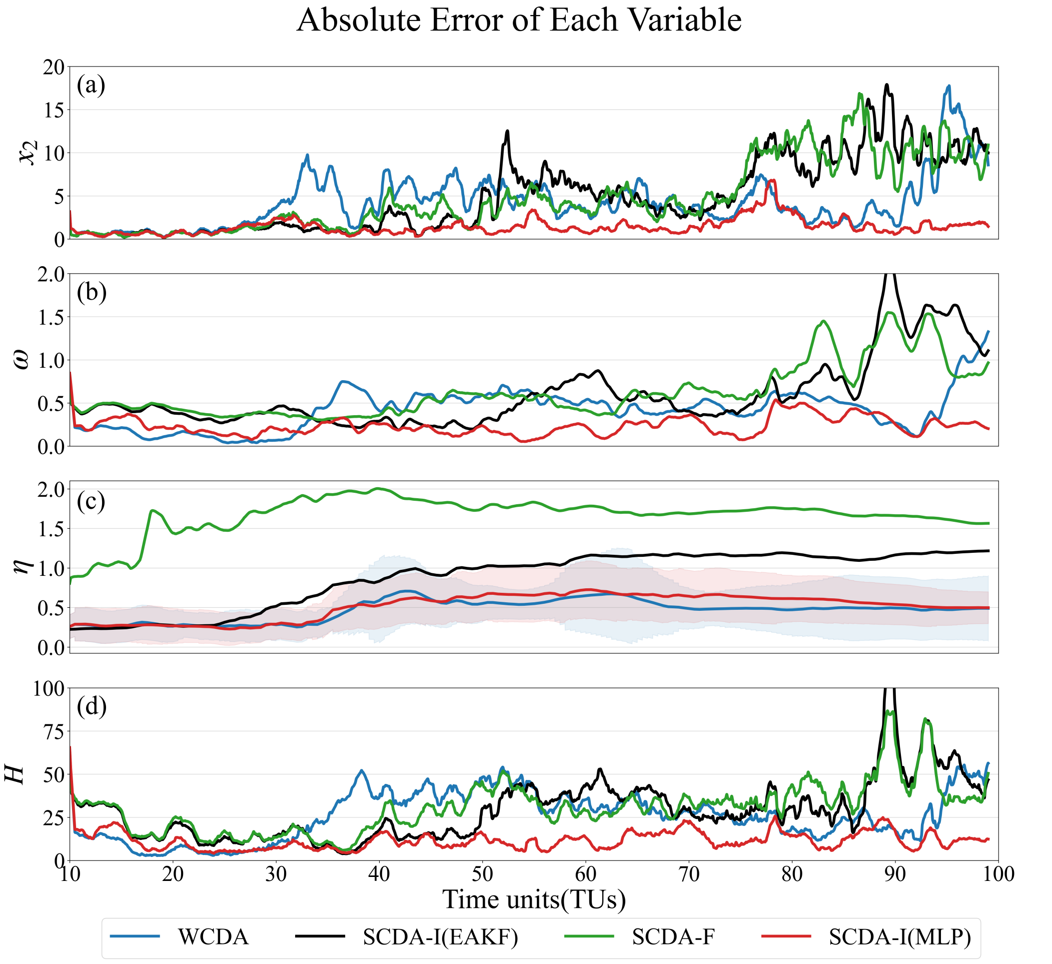

Figure 11. The absolute error of x_2 (**a**), ω (**b**), η (**c**) and H (**d**) of size 10 are collected in WCDA (blue), SCDA-I(EAKF) (black), SCDA-F (green) and SCDA-I(MLP) (red), respectively. The shaded areas in (**c**) represent the mean RMSE of 10 replicate experiments plus/minus the standard deviation of them.

For more detailed quantitative results, Table 1 presents the time-averaged RMSEs of x_2, ω, η, and φ from the experimental results with varying ensemble sizes. In these experiments, the SCDA-I experiments used the conventional EAKF and three DNN models. Examination of the data reveals clear advantages of DNN–EAKF, including reduced analysis errors and improved ACC. Notably, among the three DNN models, SLP performs best for atmospheric variables, while MLP excels for ocean and related variables. Table 1 also illustrates the relative error reduction rate and relative ACC increase rate of SCDA-I (MLP) compared to WCDA. The improvement resulting from strongly coupled data assimilation is particularly pronounced with a smaller number of ensemble members. It is worth noting that although SCDA-I using DNN–EAKF did not reduce the RMSEs of pycnocline variable η compared to WCDA (especially when N = 20 or 50), the ACCs still increased. Due to the lack of observations on η, the accuracy of η analysis is poor. The improvement of η is mainly achieved through model integration; therefore, it has a good correlation.

Table 1. The time-averaged root mean square errors (RMSE) and anomaly correlation coefficients (ACC) of x_2, ω, η, and φ under different ensemble member conditions when the observation interval of atmospheric variable is 0.1 TUs, and the ocean and sea ice is 0.5 TUs .

	RMSE											
	N = 10				**N = 20**				**N = 50**			
	x_2	ω	η	φ	x_2	ω	η	φ	x_2	ω	η	φ
WCDA	8.64	0.72	0.52	0.21	7.63	0.67	0.46	0.19	5.60	0.33	0.30	0.15
SCDA-I	10.01	0.82	0.89	0.28	9.86	0.65	0.87	0.25	7.21	0.65	1.02	0.26
SCDA-F	9.61	0.89	1.96	0.33	10.37	0.85	1.26	0.26	7.50	0.64	1.30	0.24
SCDA-I(MLP)	2.35	0.28	0.51	0.13	2.04	0.26	0.54	0.11	2.10	0.24	0.46	0.11
SCDA-I(SLP)	1.81	0.31	0.40	0.14	1.63	0.30	0.55	0.13	1.52	0.28	0.70	0.13
SCDA-I(SIP)	2.33	0.31	0.53	0.14	2.31	0.30	0.53	0.13	1.68	0.29	0.56	0.13
reduction rate	72.78%	61.50%	1.01%	39.46%	73.30%	61.43%	−16.42%	39.88%	62.52%	27.42%	−55.21%	20.4%
	ACC											
	x_2	ω	η	φ	x_2	ω	η	φ	x_2	ω	η	φ
WCDA	0.86	0.74	0.79	0.69	0.89	0.85	0.70	0.77	0.93	0.93	0.82	0.83
SCDA-I	0.81	0.76	0.84	0.59	0.84	0.84	0.91	0.66	0.90	0.85	0.90	0.64
SCDA-F	0.81	0.75	0.38	0.51	0.80	0.72	0.28	0.57	0.89	0.82	0.47	0.65
SCDA-I(MLP)	0.98	0.99	0.93	0.90	0.99	0.99	0.95	0.92	0.99	0.99	0.95	0.92
SCDA-I(SLP)	0.98	0.98	0.90	0.89	0.99	0.99	0.95	0.90	1.00	0.99	0.93	0.9
SCDA-I(SIP)	0.99	0.99	0.96	0.88	0.99	0.98	0.95	0.90	0.99	0.99	0.95	0.90
growth rate	14.07%	32.89%	16.78%	30.67%	12.04%	16.40%	34.98%	20.17%	6.33%	5.87%	16.72%	11.44%

5. Conclusions

In recent years, the research of CDA has received extensive attention, among which SCDA is considered the theoretically optimal coupled data assimilation method for reanalysis and prediction initialization. One of the key challenges of SCDA is to accurately estimate its coupled error covariance matrix, especially the cross-component error covariance. Numerous studies have shown that in ensemble-based data assimilation algorithms, the accuracy of covariance estimation is highly dependent on the ensemble size. This is because traditional ensemble-based data assimilation methods, such as EAKF, adopt linear and Gaussian assumptions, making it difficult to handle cross-component updates in SCDA.

To solve the difficulties faced by EAKF in SCDA, this paper proposes a DNN–EAKF algorithm, which incorporates machine learning and EAKF. The new algorithm employs nonlinear intra-variable relationships established by DNN during the analysis stage of the EAKF, which nonlinearly projects observation increments into the state variable space. It means that this method relies less on the Gaussian linear assumption and has the potential to handle nonlinear SCDA situations.

This study verifies the algorithm's performance using a conceptual coupled model consisting of atmospheric, oceanic, pycnocline, and sea ice variables at different spatiotemporal scales. Twin experiments are conducted by creating synthetic atmospheric, oceanic, and sea ice observations and comparing the assimilation performance of different methods. The emphasis is placed on the strong/weak CDA of atmospheric observations, and the ocean and sea ice observations are assimilated using WCDA.

Three DNN models are established based on different considerations, namely SIP, SLP, and MLP models, which are obtained to use atmosphere variables to predict the ocean. The DNN–EAKF obtained by combining these models with EAKF is used for SCDA using atmospheric observations to update ocean variables (we call it SCDA-interface or SCDA-I). The results show that DNN–EAKF performs much better in SCDA-I than the conventional EAKF method using finite ensemble members, indicating that the DNN–EAKF can better estimate the cross-component error covariances, thus providing more accurate analysis. The SCDA-I can be even better than WCDA using EAKF. This indicates that even with very few ensemble members, SCDA-I using DNN–EAKF can still provide effective information to other model components, implying its effectiveness. From Figures 8 and 9, it concludes that the RMSEs are reduced from WCDA up to 70% while the ACCs are increased up to 20% when the atmospheric observations are used to update the ocean variables directly. Figure 10 and Table 1 show that the other model components can also be improved through SCDA.

The experimental results show that DNN–EAKF has great potential to improve the ensemble-based data assimilation performance in coupled modes. However, this study primarily presents the rationale behind the DNN–EAKF algorithm and validates the concept using a relatively low-order model. Although the simple model demonstrates promising results, various challenges persist in its application to realistic high-resolution models. Specifically, the computational cost of training the machine-learning model escalates in complicated coupled models, and predictions generated during assimilation by the machine-learning model introduce computational overhead. Moreover, its superiority over EAKF in weak nonlinearity cases is not significant. In future research, our focus will be on further optimizing DNN–EAKF in additional low-order models and its application to real operational prediction models.

Author Contributions: Conceptualization, Z.S.; methodology, R.W. and Z.S.; software, R.W. and Z.S.; validation, R.W.; writing—original draft preparation, R.W.; writing—review and editing, Z.S.; visualization, R.W.; supervision, Z.S.; project administration, Z.S.; funding acquisition, Z.S. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by grants from the National Natural Science Foundation of China (42176003), the Jiangsu Provincial Innovation and Entrepreneurship Doctor Program (JSSCBS20210252).

Data Availability Statement: The data that support the findings of this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fujii, Y.; Nakaegawa, T.; Matsumoto, S.; Yasuda, T.; Yamanaka, G.; Kamachi, M. Coupled climate simulation by constraining ocean fields in a coupled model with ocean data. *J. Clim.* **2009**, *22*, 5541–5557. [CrossRef]
2. Saha, S.; Nadiga, S.; Thiaw, C.; Wang, J.; Wang, W.; Zhang, Q.; Van den Dool, H.; Pan, H.L.; Moorthi, S.; Behringer, D.; et al. The NCEP climate forecast system. *J. Clim.* **2006**, *19*, 3483–3517. [CrossRef]
3. Zhang, R.; Delworth, T.L. Impact of the Atlantic multidecadal oscillation on North Pacific climate variability. *Geophys. Res. Lett.* **2007**, *34*, 2162. [CrossRef]
4. Zhang, S.; Liu, Z.; Zhang, X.; Wu, X.; Deng, X. Coupled data assimilation and parameter estimation in coupled ocean–atmosphere models: A review. *Clim. Dyn.* **2020**, *54*, 5127–5144. [CrossRef]
5. Penny, S.G.; Bach, E.; Bhargava, K.; Chang, C.; Da, C.; Sun, L.; Yoshida, T. Strongly Coupled Data Assimilation in Multiscale Media: Experiments Using a Quasi-Geostrophic Coupled Model. *J. Adv. Model. Earth Syst.* **2019**. *11*, 1803–1829. [CrossRef]
6. Han, G.; Wu, X.; Zhang, S.; Liu, Z.; Li, W. Error covariance estimation for coupled data assimilation using a Lorenz atmosphere and a simple pycnocline ocean model. *J. Clim.* **2013**, *26*, 10218–10231. [CrossRef]
7. Lu, F.; Liu, Z.; Zhang, S.; Liu, Y. Strongly coupled data assimilation using leading averaged coupled covariance (LACC). Part I: Simple model study. *Mon. Weather Rev.* **2015**, *143*, 3823–3837. [CrossRef]
8. Smith, P.J.; Lawless, A.S.; Nichols, N.K. Treating sample covariances for use in strongly coupled atmosphere-ocean data assimilation. *Geophys. Res. Lett.* **2018**, *45*, 445–454. [CrossRef]
9. Frolov, S.; Bishop, C.H.; Holt, T.; Cummings, J.; Kuhl, D. Facilitating strongly coupled ocean–atmosphere data assimilation with an interface solver. *Mon. Weather Rev.* **2016**, *144*, 3–20. [CrossRef]
10. Yoshida, T. Covariance Localization in Strongly Coupled Data Assimilation. Ph.D. Thesis, University of Maryland, College Park, MD, USA, 2019.

11. Shen, Z.; Tang, Y.; Li, X.; Gao, Y. On the Localization in Strongly Coupled Ensemble Data Assimilation Using a Two-Scale Lorenz Model. *Earth Space Sci.* **2021**, *8*, e2020EA001465. [CrossRef]
12. Cheng, S.; Quilodrán-Casas, C.; Ouala, S.; Farchi, A.; Liu, C.; Tandeo, P.; Fablet, R.; Lucor, D.; Iooss, B.; Brajard, J.; et al. Machine Learning With Data Assimilation and Uncertainty Quantification for Dynamical Systems: A Review. *IEEE/CAA J. Autom. Sin.* **2023**, *10*, 1361–1387. [CrossRef]
13. Arcucci, R.; Zhu, J.; Hu, S.; Guo, Y.K. Deep data assimilation: Integrating deep learning with data assimilation. *Appl. Sci.* **2021**, *11*, 1114. [CrossRef]
14. Farchi, A.; Bocquet, M.; Laloyaux, P.; Bonavita, M.; Malartic, Q. A comparison of combined data assimilation and machine learning methods for offline and online model error correction. *J. Comput. Sci.* **2021**, *55*, 101468. [CrossRef]
15. Li, X.; Xiao, C.; Cheng, A.; Lin, H. Joint Estimation of Parameter and State with Hybrid Data Assimilation and Machine Learning. 2022. Available online: https://www.authorea.com/doi/full/10.22541/au.164605938.86704099 (accessed on 27 December 2023).
16. Legler, S.; Janjić, T. Combining data assimilation and machine learning to estimate parameters of a convective-scale model. *Q. J. R. Meteorol. Soc.* **2022**, *148*, 860–874. [CrossRef]
17. Brajard, J.; Carrassi, A.; Bocquet, M.; Bertino, L. 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.* **2020**, *44*, 101171. [CrossRef]
18. Vlachas, P.R.; Byeon, W.; Wan, Z.Y.; Sapsis, T.P.; Koumoutsakos, P. Data-driven forecasting of high-dimensional chaotic systems with long short-term memory networks. *Proc. R. Soc. A Math. Phys. Eng. Sci.* **2018**, *474*, 20170844. [CrossRef] [PubMed]
19. Bocquet, M.; Brajard, J.; Carrassi, A.; Bertino, L. Data assimilation as a learning tool to infer ordinary differential equation representations of dynamical models. *Nonlinear Process. Geophys.* **2019**, *26*, 143–162. [CrossRef]
20. Grooms, I. Analog ensemble data assimilation and a method for constructing analogs with variational autoencoders. *Q. J. R. Meteorol. Soc.* **2021**, *147*, 139–149. [CrossRef]
21. Pawar, S.; Ahmed, S.E.; San, O.; Rasheed, A.; Navon, I.M. Long short-term memory embedded nudging schemes for nonlinear data assimilation of geophysical flows. *Phys. Fluids* **2020**, *32*, 076606. [CrossRef]
22. Evensen, G. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res. Ocean.* **1994**, *99*, 10143–10162. [CrossRef]
23. Kalman, R.E. A new approach to linear filtering and prediction problems. *J. Basic Eng.* **1960**, *82*, 35–45. [CrossRef]
24. Luo, X.; Hoteit, I. Ensemble Kalman filtering with a divided state-space strategy for coupled data assimilation problems. *Mon. Weather Rev.* **2014**, *142*, 4542–4558. [CrossRef]
25. Anderson, J.L. An ensemble adjustment Kalman filter for data assimilation. *Mon. Weather Rev.* **2001**, *129*, 2884–2903. [CrossRef]
26. Whitaker, J.S.; Hamill, T.M. Ensemble data assimilation without perturbed observations. *Mon. Weather Rev.* **2002**, *130*, 1913–1924. [CrossRef]
27. Bishop, C.H.; Etherton, B.J.; Majumdar, S.J. Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. *Mon. Weather Rev.* **2001**, *129*, 420–436. [CrossRef]
28. Anderson, J.L. A local least squares framework for ensemble filtering. *Mon. Weather Rev.* **2003**, *131*, 634–642. [CrossRef]
29. Han, G.J.; Zhang, X.F.; Zhang, S.; Wu, X.R.; Liu, Z. Mitigation of coupled model biases induced by dynamical core misfitting through parameter optimization: Simulation with a simple pycnocline prediction model. *Nonlinear Process. Geophys.* **2014**, *21*, 357–366. [CrossRef]
30. Zhang, S. A study of impacts of coupled model initial shocks and state–parameter optimization on climate predictions using a simple pycnocline prediction model. *J. Clim.* **2011**, *24*, 6210–6226. [CrossRef]
31. Zhang, S. Impact of observation-optimized model parameters on decadal predictions: Simulation with a simple pycnocline prediction model. *Geophys. Res. Lett.* **2011**, *38*, L02702. [CrossRef]
32. Zhang, S.; Liu, Z.; Rosati, A.; Delworth, T. A study of enhancive parameter correction with coupled data assimilation for climate estimation and prediction using a simple coupled model. *Tellus A Dyn. Meteorol. Oceanogr.* **2012**, *64*, 10963. [CrossRef]
33. Zhang, S.; Winton, M.; Rosati, A.; Delworth, T.; Huang, B. Impact of enthalpy-based ensemble filtering sea ice data assimilation on decadal predictions: Simulation with a conceptual pycnocline prediction model. *J. Clim.* **2013**, *26*, 2368–2378. [CrossRef]
34. Lorenz, E.N. Deterministic nonperiodic flow. *J. Atmos. Sci.* **1963**, *20*, 130–141. [CrossRef]
35. Sluka, T.; Penny, S.; Kalnay, E.; Miyoshi, T. Strongly coupled enkf data assimilation in coupled ocean-atmosphere models. In Proceedings of the 96th AMS Annual Meeting, "Earth System Science in Service to Society", New Orleans, LA, USA, 10–14 January 2016; pp. 10–14.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.