Deep Learning Augmented Data Assimilation: Reconstructing Missing Information with Convolutional Autoencoders

YUEYA WANG,[a] XIAOMING SHI,[a] LILI LEI,[b] AND JIMMY CHI-HUNG FUNG[a,c]

[a] *Division of Environment and Sustainability, Hong Kong University of Science and Technology, Hong Kong, China*
[b] *School of Atmospheric Sciences, Nanjing University, Nanjing, China*
[c] *Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong, China*

(Manuscript received 26 October 2021, in final form 14 April 2022)

ABSTRACT: Remote sensing data play a critical role in improving numerical weather prediction (NWP). However, the physical principles of radiation dictate that data voids frequently exist in physical space (e.g., subcloud area for satellite infrared radiance or no-precipitation region for radar reflectivity). Such data gaps impair the accuracy of initial conditions derived from data assimilation (DA), which has a negative impact on NWP. We use the barotropic vorticity equation to demonstrate the potential of deep learning augmented data assimilation (DDA), which involves reconstructing spatially complete pseudo-observation fields from incomplete observations and using them for DA. By training a convolutional autoencoder (CAE) with a long simulation at a coarse "forecast" resolution (T63), we obtained a deep learning approximation of the "reconstruction operator," which maps spatially incomplete observations to a model state with full spatial coverage and resolution. The CAE was applied to an incomplete streamfunction observation (~30% missing) from a high-resolution benchmark simulation and demonstrated satisfactory reconstruction performance, even when only very sparse (1/16 of T63 grid density) observations were used as input. When only spatially incomplete observations are used, the analysis fields obtained from ensemble square root filter (EnSRF) assimilation exhibit significant error. However, in DDA, when EnSRF takes in the combined data from the incomplete observations and CAE reconstruction, analysis error reduces significantly. Such gains are more pronounced with sparse observation and small ensemble size because the DDA analysis is much less sensitive to observation density and ensemble size than the conventional DA analysis, which is based solely on incomplete observations.

SIGNIFICANCE STATEMENT: Data assimilation plays a critical role in improving the skills of modern numerical weather prediction by establishing accurate initial conditions. However, unobservable regions are common in observation data, particularly those derived from remote sensing. The nonlinear relationship between data from observable regions and the physical state of unobservable regions may impede DA efficiency. As a result, we propose that deep learning be used to improve data assimilation in such cases by reconstructing a spatially complete first guess of the physical state with deep learning and then applying data assimilation to the reconstructed field. Such deep learning augmentation is found effective in improving the accuracy of data assimilation, especially for sparse observation and small ensemble size.

KEYWORDS: Data assimilation; Numerical weather prediction/forecasting; Machine learning

1. Introduction

The assimilation of satellite and radar remote sensing data has greatly enhanced numerical weather prediction (NWP) by improving initial conditions (Alley et al. 2019; Jung et al. 2008; Simmons and Hollingsworth 2002). However, due to the nature of remote sensing techniques, significant data gaps exist (e.g., Zheng et al. 2021). Clouds and heavy precipitation, for example, can significantly corrupt satellite infrared radiance data; thus, most weather centers only assimilate radiance in cloud-free regions, though some have begun to include the information above cloud tops (Geer et al. 2018). In the case of radar remote sensing, the radar echoes of precipitation are clustered and provide limited innovations in regions out of the rainfall areas (Sodhi and Fabry 2020). As velocity observations can only be derived from the areas with radar echoes, the improvement of the wind analysis field is also limited. Furthermore, these areas may be obscured by high-level clouds, limiting the improvement from radiance observations of spaceborne infrared images (Errico et al. 2007). As a result, without the key information outside of the observation range, we cannot obtain an analysis field with a completely consistent dynamics and thermodynamics environment (Fabry and Meunier 2020). Some multiscale approaches have been proposed to smooth either the background error covariance or observations to allow the use of a larger localization window and increase the impact region of observation (Caron and Buehner 2018; Miyoshi and Kondo 2013; Ying 2020).

In this study, we adopt a different path by proposing the use of deep learning to fill the data voids in remote sensing scenarios as an enhancement to conventional data assimilation (DA). The convolutional autoencoder (CAE) is a highly effective deep learning model structure that has been widely used in image denoising and inpainting (Mao et al. 2016; Xie et al. 2012). It represents high-dimensional complex data using an unsupervised

Corresponding author: Xiaoming Shi, shixm@ust.hk

DOI: 10.1175/MWR-D-21-0288.1

low-dimensional latent space. We show, using a prototype problem, that it is possible to use a CAE to reconstruct full dynamical fields from limited observations. The reconstructed dynamical field then serves as pseudo-observations for DA and enhances its impact. We term this approach as deep learning augmented data assimilation (DDA).

The principle idea of DDA can be explained as follows. A DA cycle, in general, has two steps: *prediction* and *analysis*. The prediction step maps $P(\mathbf{x}_j|\mathbf{y}_j)$ to $P(\mathbf{x}_{j+1}|\mathbf{y}_j)$, where $\mathbf{x}_j$ is the atmospheric state vector at time j, $\mathbf{y}_j$ is the observation vector, and P denotes probability distribution functions. This step is completed by integrating an NWP model in time. The analysis step computes $P(\mathbf{x}_{j+1}|\mathbf{y}_j) \mapsto P(\mathbf{x}_{j+1}|\mathbf{y}_{j+1})$, after obtaining the prediction $P(\mathbf{x}_{j+1}|\mathbf{y}_j)$ and new observation $\mathbf{y}_{j+1}$, by the application of Bayes's rule. For the reasons stated above, observation $\mathbf{y}$ may have different spatial coverage than state variables $\mathbf{x}$. We want to find a way to map $\mathbf{y}$ to a newly reconstructed pseudo-observation $\mathbf{x}^r$ that covers the entire domain of the NWP model[1] to fill in the data gaps. Deep learning is a promising solution to this difficult task. In the usual DA, the observation is related to a state vector by an observation operator, i.e., $\mathbf{y} = h(\mathbf{x})$. What we propose here is using neural networks to approximate h^{-1}, with which we can compute $\mathbf{y} \mapsto \mathbf{x}^r$, and thus converting the analysis step of DA to computing $P(\mathbf{x}_{j+1}|\mathbf{x}^r_{j+1})$, instead of $P(\mathbf{x}_{j+1}|\mathbf{y}_{j+1})$.

Training a deep learning model to approximate h^{-1} necessitates a large number of samples that reveal the relationship between atmospheric states and spatially incomplete observations. In practice, such samples can be obtained from historical ensemble simulations, which contain a massive amount of data and might be adequate for training deep learning models to achieve high fidelity, although evaluation is needed for any given application. However, training machine learning models for real NWP cases is computationally expensive and time-consuming. This study uses a simple barotropic vorticity model as a proof-of-concept for the method before future applications to realistic NWP cases. As a result, for DA and forecast experiments, we employ a barotropic vorticity equation model. The impact of introducing a deep learning h^{-1} approximation in DA is evaluated and demonstrated. This h^{-1} approximation is referred to as the "reconstruction operator" in this context.

Previous studies have applied machine learning techniques to observation bias correction in DA (Jin et al. 2019) and reduced-order deep DA (Casas et al. 2020). The current research takes a novel approach to improve DA. Because our goal is not to retrieve a variable from collocated remote sensing data, our approach differs significantly from traditional remote sensing retrieval (Aires et al. 2002; Bobylev et al. 2009). Instead, we are trying to "generate" data about unobservable regions. In the context of NWP, this generation should be conditioned on observation information.

2. Models and methods

a. Barotropic vorticity equation (BVE)

The idealized atmospheric model we used is the spectral BVE model on a sphere provided by the Geophysical Fluid Dynamics Laboratory (GFDL). Following Vallis et al. (2004), the governing equation of the BVE model is

$$\frac{\partial \zeta}{\partial t} + J(\psi, \zeta + f) = S - r\zeta + \kappa\nabla^4\zeta, \tag{1}$$

where ψ is the streamfunction, ζ is the barotropic vorticity, f is the Coriolis parameter, and J is the Jacobian operator. On the right-hand side are stochastic stirring, linear damping, and hyper-diffusion, respectively. Details on how these terms were implemented are documented in Vallis et al. (2004), who demonstrated that this simple model can qualitatively reproduce extratropical circulation variability at large scales. A Markov process with a decorrelation time scale of 2 days represents the stirring term S, which represents the effect of baroclinic eddies on barotropic flow. It stirs a small range of wavenumbers in the spectral space and its effect in physical space is limited to Northern Hemisphere (NH) midlatitudes.

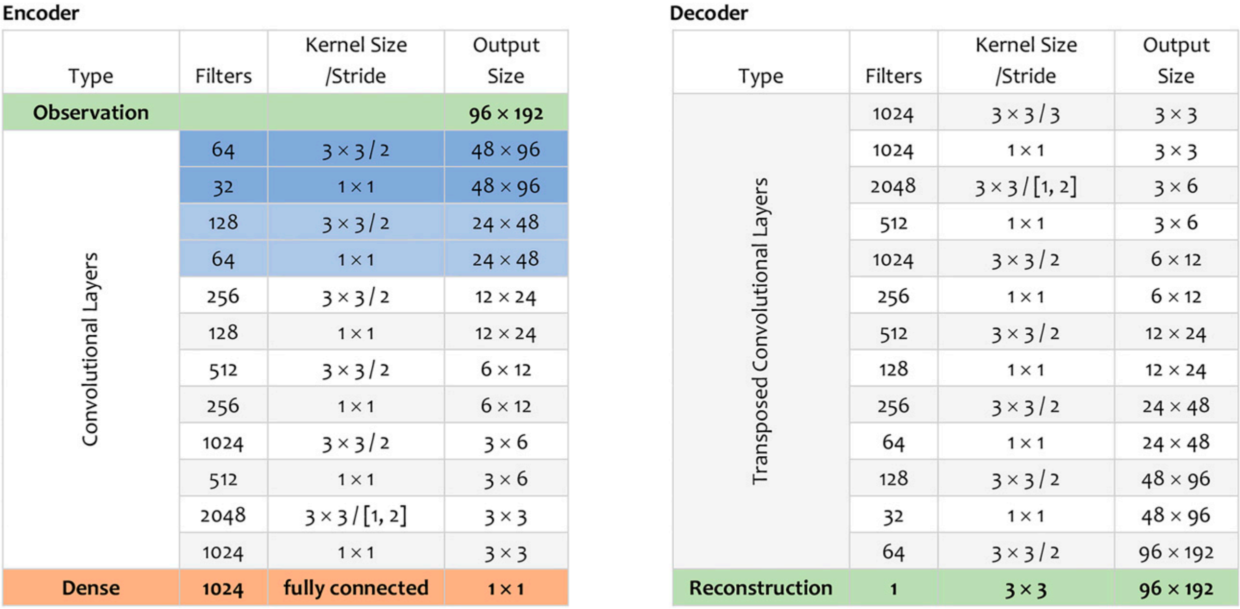
b. Convolutional autoencoder

CAEs are a class of methods in deep learning and have been extensively used in computer vision. The dimension of $\mathbf{z}$ is usually much smaller than that of $\mathbf{x}$. As a result, the encoding process is a type of compression. The decoder, which is implemented as a series of transposed convolutional layers, computes the decompression mapping, $\mathbf{z} \mapsto \mathbf{x}^r$, where $\mathbf{x}^r$ has the same dimension as $\mathbf{x}$ and is a reconstruction of the original state vector (image). A CAE is trained by minimizing a loss function, such as the mean squared error (MSE) of $\mathbf{x}^r$ compared to $\mathbf{x}$, using stochastic gradient descent, which optimizes convolutional and transposed convolutional layer parameters. If we think of the training dataset for a CAE as an ensemble, which implicitly describes a probability distribution of $\mathbf{x}$, then, in parallel to DA, we can describe the two steps of a CAE as *encoding*, which computes $P(\mathbf{x}) \mapsto P(\mathbf{z}|\mathbf{x})$, and *decoding*, which computes $P(\mathbf{z}|\mathbf{x}) \mapsto P(\mathbf{x}^r|\mathbf{z})$.

Because our intention here is to reconstruct a dynamical variable field from observations, we need to change the CAE from its usual configuration. The input to our encoder should be the observation $\mathbf{y} = h(\mathbf{x})$, while the output of the decoder is still $\mathbf{x}^r$. In the loss function, we compare $\mathbf{x}^r$ against the dynamical field $\mathbf{x}$ that generates $\mathbf{y}$. In other words, for our reconstruction operator, the encoder part computes $P(\mathbf{y}) \mapsto \mathrm{P}(\mathbf{z}|\mathbf{y})$, and the decoder part computes $P(\mathbf{z}|\mathbf{y}) \mapsto P(\mathbf{x}^r|\mathbf{z})$.

The structure of the CAE we used is shown in Fig. 1. Our baseline group of experiments have a dense observation grid, which means every state variable could be observed. For the baseline group, the encoder has 12 convolutional layers and maps an observation field (spatially incomplete streamfunction) of the size 96 × 192 to a latent vector of the length 1024. The

[1] The symbol $\mathbf{x}^r$ is used here, instead of $\mathbf{y}^r$, because in the context of remote sensing, $\mathbf{y}$ may be impossible to have a full spatial coverage (e.g., reflectivity is unavailable in clear-sky regions). This implies a conversion from remote sensing to model state variables. The prototype problem we studied here did not involve such conversion, but CAE can be structured to perform such tasks.



Encoder

Type	Filters	Kernel Size /Stride	Output Size
Observation			**96 × 192**
Convolutional Layers	64	3 × 3 / 2	48 × 96
	32	1 × 1	48 × 96
	128	3 × 3 / 2	24 × 48
	64	1 × 1	24 × 48
	256	3 × 3 / 2	12 × 24
	128	1 × 1	12 × 24
	512	3 × 3 / 2	6 × 12
	256	1 × 1	6 × 12
	1024	3 × 3 / 2	3 × 6
	512	1 × 1	3 × 6
	2048	3 × 3 / [1, 2]	3 × 3
	1024	1 × 1	3 × 3
Dense	**1024**	**fully connected**	**1 × 1**

Decoder

Type	Filters	Kernel Size /Stride	Output Size
Transposed Convolutional Layers	1024	3 × 3 / 3	3 × 3
	1024	1 × 1	3 × 3
	2048	3 × 3 / [1, 2]	3 × 6
	512	1 × 1	3 × 6
	1024	3 × 3 / 2	6 × 12
	256	1 × 1	6 × 12
	512	3 × 3 / 2	12 × 24
	128	1 × 1	12 × 24
	256	3 × 3 / 2	24 × 48
	64	1 × 1	24 × 48
	128	3 × 3 / 2	48 × 96
	32	1 × 1	48 × 96
	64	3 × 3 / 2	96 × 192
Reconstruction	**1**	**3 × 3**	**96 × 192**

FIG. 1. The structure of the convolutional autoencoder used in this study. The full encoder accepts observation images ($\mathbf{y}$) of the size 96 × 192 and has 12 convolutional layers and 1 fully connected layer. The modified encoder for experiments with sparse and very sparse observations removes the first 2 and 4 layers, respectively, and consequently has 10 or 8 convolutional layers (indicated by a blue shading). The latent space has a dimension of 1024. There are 14 transposed convolution layers in the decoder. The reconstructed decoder output ($\mathbf{x}^r$) has the same dimension as the full input observation, 96 × 192. Each convolutional layer or transposed convolution layer is followed by a rectified linear unit as the activation function, except that the last layers of encoder and decoder (last row of each table) have no activation functions.

decoder has 14 transposed convolutional layers and maps the latent vector to a reconstructed field of size 96 × 192. For groups of experiments with a sparse observation grid of 48 × 96 and a very sparse observation grid of 24 × 48, we remove the first two and four convolutional layers in the encoder, respectively, to make it compatible with coarse resolution data. To get a reconstructed streamfunction field with full coverage of the size 96 × 192, the structure of the decoder part is the same for different groups of experiments. In this CAE, the layers with the 1 × 1 filters work as compression layers which reduce the number of channels from the previous layer.

The loss function we used for training the CAE is

$$\mathscr{L} = \frac{1}{2}(\mathscr{L}_{\mathrm{MSE}} + \mathscr{L}_{\mathrm{MS\text{-}SSIM}}), \tag{2}$$

where $\mathscr{L}_{\mathrm{MSE}}$ is the MSE, and $\mathscr{L}_{\mathrm{MS\text{-}SSIM}}$ is the multiscale structural similarity index measure (MS-SSIM) loss. The $\mathscr{L}_{\mathrm{MSE}}$ is a measure of absolute error and is calculated with normalized streamfunction fields (by mean and standard deviation). The MS-SSIM ($I_{\mathrm{MS\text{-}SSIM}}$) measures the structural similarity between two images, and its value is between zero and one, with a higher index value indicating higher similarity as shown in Eq. (7) of Wang et al. (2003).

The approach is based on modeling image luminance, contrast, and structure at multiple scales. The overall MS-SSIM evaluation is obtained by combining the measurement at different scales. The MS-SSIM has resulted in much better performance than the single-scale SSIM approach but at the cost of a relatively lower processing speed. Thus, $\mathscr{L}_{\mathrm{MS\text{-}SSIM}}$ for a single pair of images is defined as $(1 - I_{\mathrm{MS\text{-}SSIM}})$.

c. *Ensemble square root filter (EnSRF)*

The serial EnSRF method (Tippett et al. 2003; Whitaker and Hamill 2002) is a deterministic ensemble filter formulation and used in our study. The flow-dependent representation of background error covariance is provided by an ensemble of model state realizations. Because the observations are not perturbed, the EnSRF method does not introduce additional sampling noise like the classical ensemble Kalman filter (EnKF) (Burgers et al. 1998). The update equations of EnSRF can be written as

$$\overline{\mathbf{x}}^a = \overline{\mathbf{x}}^b + \mathbf{K}[\mathbf{y} - h(\overline{\mathbf{x}}^b)], \tag{3}$$

$$\mathbf{x}_i'^a = \beta(\mathbf{I} - \alpha\mathbf{K}\mathbf{H})\mathbf{x}_i'^b, \tag{4}$$

where an overbar denotes the ensemble mean, a prime denotes the perturbation of an ensemble member, and subscript i the ith member among the N ensemble members. The superscripts a and b in the equations denote the analysis and the background states, respectively. The α is the square root modification factor, which is defined as

$$\alpha = \left[1 + \sqrt{\mathbf{R}(\mathbf{H}\mathbf{P}^b\mathbf{H}^{\mathrm{T}} + \mathbf{R})^{-1}}\right]^{-1}, \tag{5}$$

where $\mathbf{H}$ is the linearized version of the observation operator h. The background state and observation error covariances are denoted as **P** and **R**, respectively. The matrix **K** is the Kalman gain matrix, calculated as follows:

$$\mathbf{K} = \mathbf{P}^b\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{P}^b\mathbf{H}^{\mathrm{T}} + \mathbf{R})^{-1}. \tag{6}$$

We did not calculate the **P** and **H** in the equation directly. Instead, $\mathbf{HP}^b\mathbf{H}^{\mathrm{T}}$ and $\mathbf{P}^b\mathbf{H}^{\mathrm{T}}$ are approximated based on the ensemble members through the following:

$$\mathbf{HP}^b\mathbf{H}^{\mathrm{T}} \cong \frac{1}{N-1}\sum_i[h(\mathbf{x}_i^b) - \overline{h(\mathbf{x}^b)}][h(\mathbf{x}_i^b) - \overline{h(\mathbf{x}^b)}]^{\mathrm{T}}, \tag{7}$$

$$\mathbf{P}^b\mathbf{H}^{\mathrm{T}} \cong \frac{1}{N-1}\sum_i(\mathbf{x}_i^b - \overline{\mathbf{x}^b})[h(\mathbf{x}_i^b) - \overline{h(\mathbf{x}^b)}]^{\mathrm{T}}. \tag{8}$$

The Gaspari–Cohn localization (Gaspari and Cohn 1999) was applied to reduce the sampling noise due to the use of limited-size ensembles. The distance at which covariance tapers to zero is referred to as the localization radius r, and $r = 15$ grid spacings were used in this study. To maintain adequate ensembles spread, a multiplicative inflation factor ($\beta = 1.15$) is used (Anderson and Anderson 1999). We tested different values of the localization radius (5, 10, 15 and 20) first and found $r = 15$ shows the minimum analysis error. Then three groups of inflation factors (1.05, 1.10 and 1.15) were tested at $r = 15$. The inflation factor of 1.15 yielded the minimum analysis error for the baseline group and this value was chosen for further evaluation. In addition, we found the assimilation results with complete observations are less sensitive to the choice of the inflation factor in the baseline dense group.

3. Experiments design

a. *CAE training*

A large number of simulation samples are needed to train a deep learning model for building the relationship between the full streamfunction field and incomplete observations. For the training of the CAE, we ran the GFDL model for 50 years at T63 resolution, which has a grid spacing of approximately 1.9° × 1.9° at the equator and 96 × 192 grid points in total. The 6-hourly data were saved, and the first 100 days of data were discarded as spinup. As a result, we have roughly 73 000-time slices in total. Then, different time slices from this dataset are randomly shuffled and partitioned into training and validation datasets, each of which contains 90% and 10% of the total dataset, respectively. The validation dataset is required because overfitting may occur in later stages of the model training process. Thus, we train the CAE with a training dataset and evaluate the loss over the validation set after each epoch; a new result is saved only if the validation loss is smaller than the previous epoch.

Streamfunction from the datasets is the state variable **x** here and was used to produce the spatially incomplete observation **y**. The observation **y** is the same as the streamfunction at grid points where $\zeta \le 5 \times 10^{-6}$ s^{-1} in the Northern Hemisphere and $\zeta \ge -5 \times 10^{-6}$ s^{-1} in the Southern Hemisphere. Other grid points are masked as unobservable regions. The streamfunction data were normalized with its mean and standard deviation before inputting to the CAE and stays within −10 to 10. Large constants of 100 and −100 are used to mask unobservable grid points in the Northern and Southern Hemispheres, respectively. This masking scheme is intended to reflect the extratropical association of clouds with storminess (strong vorticity). Masking occurs mostly in the NH because the Southern Hemisphere has no stirring and thus few eddies. Unobservable regions cover an average of 30.2% of the NH and 15.6% of the global area. Random noise with the amplitude of 0.05 was added to the normalized streamfunction observation to enhance the CAE's error tolerance.

Additionally, as described in the next section, for testing the efficacy of the CAE reconstruction with respect to different observation densities, we have three main groups of experiments, for which the observation grids are 96 × 192 (Dense), 48 × 96 (Sparse), and 24 × 48 (Very-Sparse), respectively. Thus, we trained three CAE models in total, and the input observation to the CAEs for the sparse and very sparse groups is coarsened accordingly.

In summary, input to the CAE is the (normalized) streamfunction covering 70% area of the NH and most of the Southern Hemisphere with artificial noise and being coarsened where necessary. The CAE output is the reconstructed streamfunction that covers the entire globe and always has a full resolution (on the 96 × 192 grid). During the CAE training, the reconstructed full field is compared to the spatially complete streamfunction field without masks and noise.

We configured and trained our CAE with TensorFlow 2.4.1. The adaptive moment estimation (Adam) algorithm was used to optimize the CAE's parameters with 160 epochs of iteration, which were carried out in four 40-epoch stages, with learning rates of 1×10^{-4}, 5×10^{-5}, 2.5×10^{-5}, and 1×10^{-5}, respectively. The loss function for the validation dataset is ~1.1×10^{-3} by the end of the training, and the MSE by the end is ~0.8×10^{-3} (for the normalized streamfunction data).

b. *Assimilation experiments*

We evaluate the performance of the conventional use of EnSRF and its combination with deep learning using the BVE model at T63 resolution. The benchmark simulation ("truth run"), on the other hand, is obtained by running the BVE model at the higher T85 resolution, which has a total of 128 × 256 grid points. This different resolution was purposefully chosen for evaluating the performance of the trained CAE because we cannot have a perfect model for generating training datasets and forecasts in real-world applications. If we use T63 resolution for the truth run also, the benefits of deep learning reconstruction become smaller when having dense observations but are still substantial if very sparse observations are used (cf. the appendix). To be consistent with the T63 simulation, we ran the T85 simulation for 51 years and then discarded the first 50 years; the data from the last year was used as the "truth."

Continuous DA cycling experiments are conducted over 1 year with a cycling period of 1 day, and performance metrics (RMSE) are averaged over the last 100 cycles. We initialized the ensemble with random instances from the set of 50-yr model states obtained

TABLE 1. Summary of the data assimilation experiments.

Group	Observation	Description	Grid
Dense Obs (baseline)	FullObs	Truth streamfunction interpolated onto T63 grid with full spatial coverage	96 × 192
	PartObs	Truth streamfunction interpolated onto T63 grid and partly masked based on vorticity thresholds	96 × 192 (masking 30% of NH)
	MixObs	Mixed field from PartObs and CAE reconstruction	96 × 192
Sparse Obs	PartObsS	1/4 of PartObs observation	48 × 96 (masking 30% of NH)
	MixObsSS	Mixed field from PartObsS and CAE reconstruction coarsened onto the sparse grid	48 × 96
	MixObsSC	Mixed field from PartObsS and complete CAE reconstructed field on the T63 grid	96 × 192
Very-Sparse Obs	PartObsVS	1/16 of PartObs observation	24 × 48 (masking 30% of NH)
	MixObsVSS	Mixed field from PartObsVS and CAE reconstruction coarsened onto the very sparse grid	24 × 48
	MixObsVSC	Mixed field from PartObsVS and complete CAE reconstructed field on the T63 grid	96 × 192

from a single continuous integration (e.g., Sakov and Oke 2008; Blyverket et al. 2019). A single analysis cycle includes a prediction step in which the ensemble is propagated forward for 1 day and a filter update step in which the prior ensemble from the prediction is fused with observations to form the posterior ensemble (analysis).

To evaluate the efficacy of deep learning reconstruction when different amounts of observation are available, we conduct three main groups of experiments, which are referred to as “Dense Observation,” “Sparse Observation,” and “Very-Sparse Observation” groups. Table 1 summarizes the salient characteristics of each group.

1) Dense Observation

This is a baseline group with the observation grid being the same as the T63 model grid (96 × 192), onto which the truth streamfunction is interpolated. This group comprises three assimilation experiments, “FullObs,” “PartObs,” and “MixObs.” FullObs experiment has observation with full spatial coverage (i.e., no masking based on vorticity). The FullObs was adopted to indicate the upper limit of assimilation accuracy if CAE reconstruction were “perfect.” With the scheme described in section 3a, the PartObs experiment has spatially incomplete observations that are masked based on vorticity. In FullObs and PartObs, random observation noise with an amplitude of 1.5×10^6 m^2 s^{-1} was added to the observation, accounting for approximately 20% of the standard deviation of the observation. The same random noise was added to the sparse and very sparse groups. MixObs generates observations by combining available observations from PartObs and the CAE reconstruction, with values from the former in the observable area and the latter in the unobservable area. The CAE reconstruction is based on PartObs.

2) Sparse Observation

In this group, the observation grid resolution is halved in each direction, so without CAE reconstruction, only 1/4 of the T63 grid points are observed. The PartObsS experiment in this group has spatially incomplete observations with vorticity-based masking. In the MixObsSS experiment, observations are obtained by combining available observations from PartObsS and CAE reconstruction coarsened onto the sparse observation grid. As a result, the observation in MixObsSS is on the same grid as the observation in PartObsS. The MixObsSC experiment, on the other hand, keeps the full-resolution CAE reconstruction and combines it with available observation in PartObsS; thus, its observation is on the T63 grid. The CAE reconstruction is based on PartObsS.

3) Very-Sparse Observation

This group has very sparse observation for evaluating the performance of using the CAE reconstruction to enhance assimilation when very limited observations can be obtained. The observation grid resolution is further reduced, so the PartObsVS experiment here has only 1/16 of the observations in PartObs. Based on the very sparse observation and the reconstructed results, we still have two types of mixed observations; the observation grids of MixObsVSS and MixObsVSC have 24 × 48 and 96 × 192 points, respectively. The CAE reconstruction is based on PartObsVS.

All the experiments described above are conducted with 80 ensemble members, which is an adequate ensemble size and seemingly affordable by operational NWP centers. However, when the ensemble size is small, the potential benefits of the new method must also be evaluated. We also ran experiments with a small ensemble of 10 members. At the end of the following section, the results of those small ensemble experiments are compared to those of larger ensembles. In this study, sensitivity experiments on the observation density and the number of ensemble members used the parameters tuned with respect to the dense observation group. The qualitative nature of some of the results might change if the localization and inflation are tuned for each unique case, which we will test in future work.

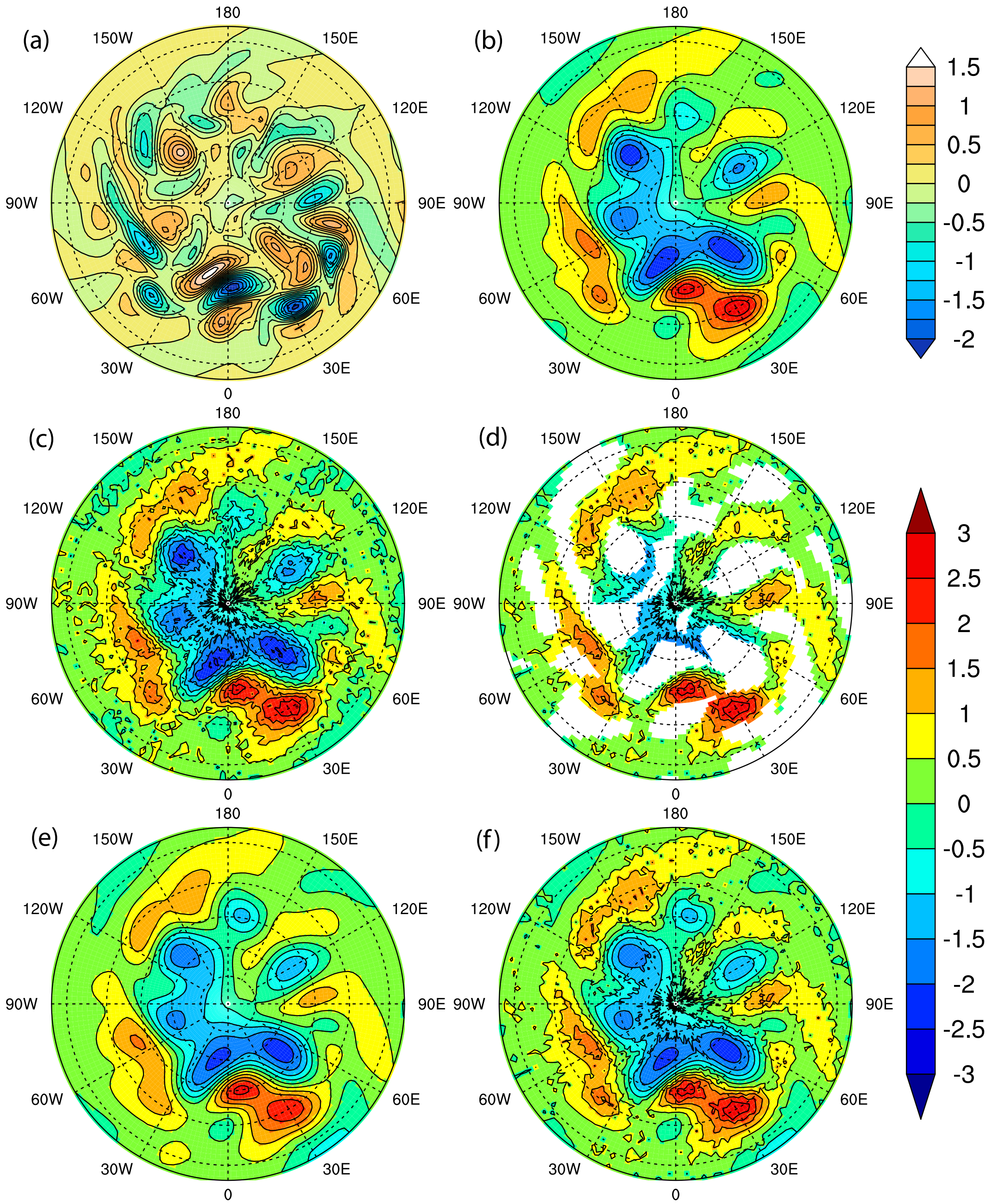


FIG. 2. The (a) vorticity (10^{-6} s^{-1}) and streamfunction (10^{6} m^{2} s^{-1}) on a randomly selected day from (b) true field, (c) FullObs observation, (d) PartObs observation, (e) CAE reconstruction based on PartObs, and (f) MixObs, which is generated by combining the CAE reconstruction in(e) and the PartObs observed streamfunction in (d). The Southern Hemisphere is not shown because there is no stochastic stirring and therefore few eddies.

4. Results

a. Deep learning

Figure 2 shows the vorticity and compares one instance of the streamfunction field from true field, observation, and CAE reconstruction. Figure 2c is the dense observation field of FullObs which was generated by adding random noise to the true field (Fig. 2b), and in PartObs (Fig. 2d), information is missing in regions where the streamfunction has relatively low values, which is usually associated with high vorticity (Fig. 2a). With limited information from the PartObs, the CAE reconstructed streamfunction (Fig. 2e) resembles the true full field closely. The CAE reconstruction contains minor errors in some details, such as the intensity of local maxima and minima, but gains accurate patterns in the large-scale distribution of eddies. It also smooths out the random noise in

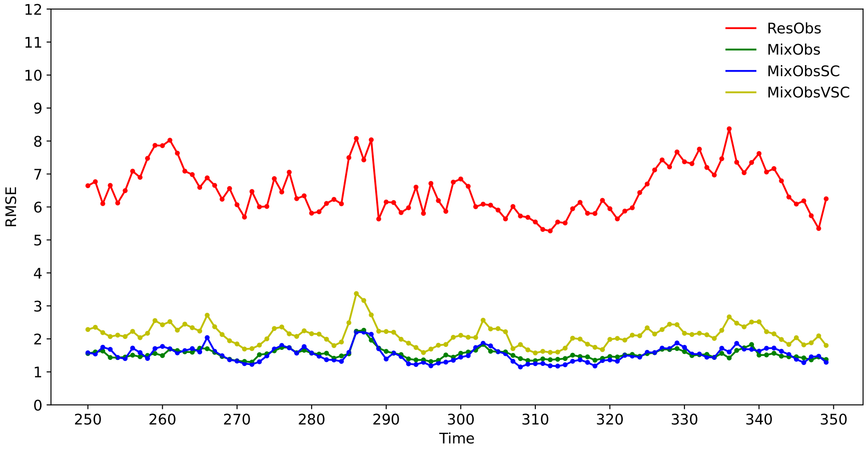


FIG. 3. The RMSE (10^6 m^2 s^{-1}) of CAE reconstructions and analog retrieval (ResObs) compared with truth observation of the last 100 days. MixObs, MixObsSC, and MixObsVSC are defined in section 3 and summarized in Table 1. ResObs denotes the analog retrieval, which is the full streamfunction field retrieved from the CAE training dataset by finding the instance that resembles the spatially incomplete observations (PartObs) most (measured by RMSE at observable grid points).

the observations, which might benefit simulations in real applications. The mixed streamfunction field (MixObs; Fig. 2f) reserves the partial observations which we can get directly from PartObs and complements the missing information with the CAE reconstruction. This mixing makes the representation in observable regions more accurate but also brings back the random noise in observable regions.

To demonstrate that the CAE indeed learned a "skillful" of reconstruction instead of simply picking out resembling snapshots from training data, we compared the root-mean-square error (RMSE) in the mixed observations (MixObs, MixObsSC, and MixObsVSC) with the analog retrieval (ResObs), in which we select the instance in the CAE training dataset that *resembles* the PartObs fields most (measured by RMSE in observable regions). Figure 3 depicts the RMSE of those four groups versus truth data (including all grid points) over the last 100 days of assimilation cycling. We can see that the RMSE of the mixed fields of CAE reconstructions and PartObs is on the order of 2 × 10^6 m^2 s^{-1}, whereas the RMSE of the analog retrievals is 4–5 times higher than the MixObsVSC. Therefore, the CAE reconstruction is highly skillful and differs from "retrieving" images from its training dataset.

Figure 3 also shows that MixObs has comparable error to MixObsSC, suggesting that the CAE model can reconstruct the streamfunction accurately even with only 1/4 of the dense observations. The mixed result of the reconstruction based on 1/16 of dense observations, on the other hand, has approximately 1.5 times the error as the other two types of mixed observations. As a result, more observations must be fed into the CAE model to reduce the CAE reconstruction error.

b. *Assimilation*

1) BASELINE ASSIMILATION

Figures 4a–c show the analyzed streamfunction for the day shown in Fig. 2. The analysis fields from the FullObs (Fig. 4a) and MixObs (Fig. 4c) are closer to the true field (Fig. 4g) than that from the PartObs (Fig. 4b). The difference between the analyzed streamfunction and the true field is larger for the PartObs (Fig. 4e) than for the FullObs (Fig. 4d) and MixObs (Fig. 4f). The analyzed results from the complete observations contain more details which are consistent with the true field. Figure 4h shows the forecast and analysis RMSE of the ensemble means of the three assimilation experiments (FullObs, PartObs, and MixObs) against the truth data for the whole cycling process in the baseline group (Dense). The large discrepancies between ensemble mean priors and posteriors are mainly caused by model errors that are due to the stochastic stirring term in the BE model and the resolution differences between the benchmark simulation and ensemble simulations. When we look at the RMSE of the posterior ensemble mean, we see that MixObs have lower analysis error than PartObs for almost all assimilation cycles. PartObs analysis has a mean last-100-day RMSE of 1.80 × 10^6 m^2 s^{-1}, which is 64% greater than MixObs analysis. Although the RMSE of MixObs analysis is still much larger than the RMSE of FullObs analysis, for which spatially complete observations are made available, the improvement due to CAE reconstruction is still quite notable in these experiments with dense observations.

Besides, the different levels of prior ensemble errors in those experiments have a significant impact on posterior ensemble errors. As shown in Fig. 4h, the PartObs forecast has the greatest error of 3.60 × 10^6 m^2 s^{-1}; in contrast, the two experiments with full-coverage observations in FullObs and MixObs have relatively lower RMSE of 2.51 × 10^6 and 2.88 × 10^6 m^2 s^{-1}, respectively. What should be mentioned is that the difference between the average prior error of the ensemble of MixObs and FullObs is larger than the difference between their posterior errors. This indicates that the error growth rate of MixObs experiments is relatively slow, potentially resulting from the denoising ability of deep learning reconstruction. These findings demonstrated that by supplementing assimilation with

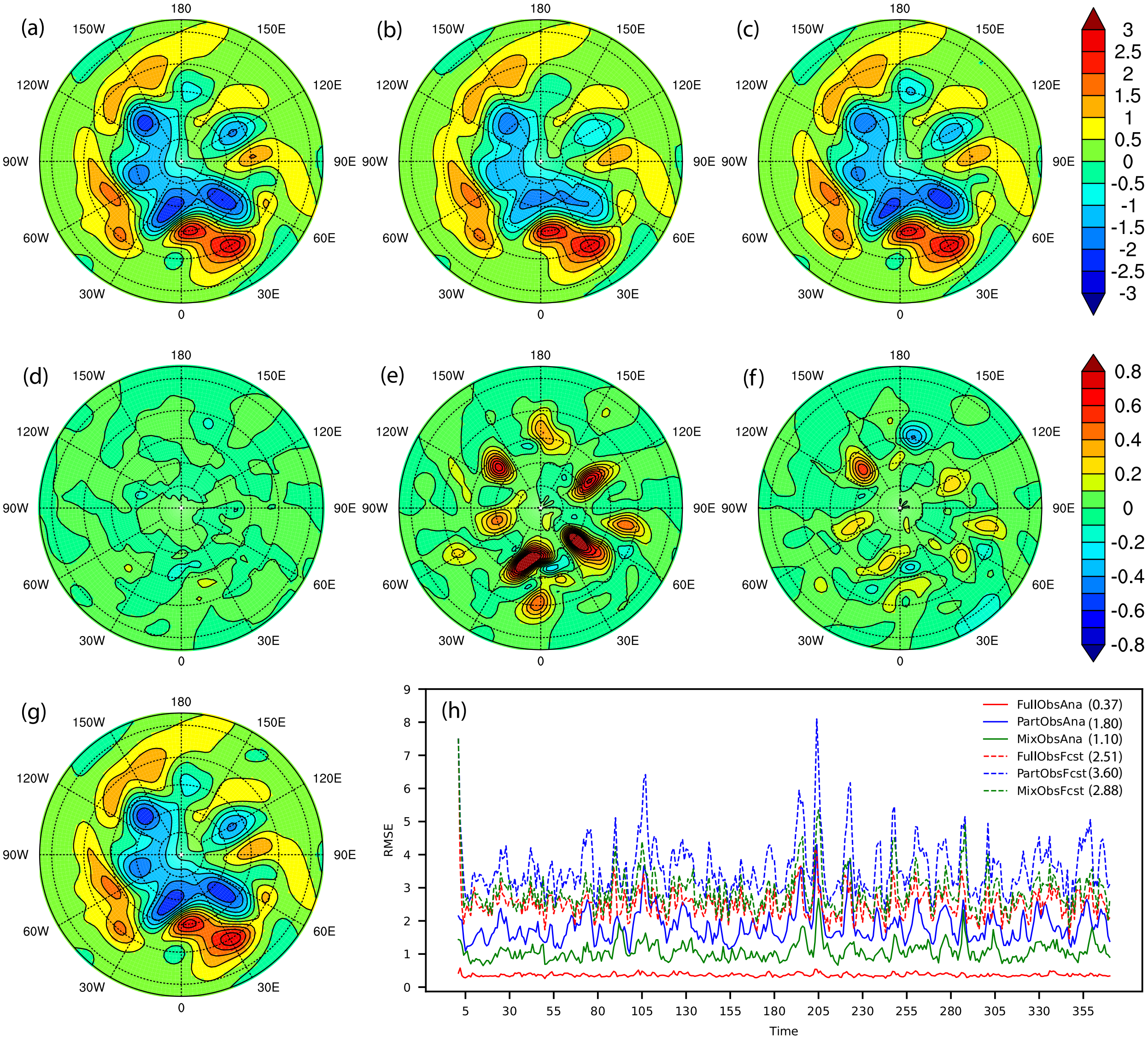


FIG. 4. Analyzed streamfunction from (a) FullObs, (b) PartObs, (c) MixObs, the difference between analyzed and (g) true streamfunction from (d) FullObs, (e) PartObs, (f) MixObs for the day shown in Fig. 2 and (h) root-mean-square error (RMSE) (10^6 m^2 s^{-1}) of the prior ensemble (forecast) mean (dashed lines) and posterior ensemble (analysis) mean (solid lines) during the assimilation cycles for the FullObs (red lines), PartObs (blue lines), and the MixObs (green lines) experiments. The average RMSE for the last 100 days is shown as the numbers in brackets in the legend. The initial (day zero) RMSE is 7.5 × 10^6 m^2 s^{-1}.

CAE reconstruction, we can obtain significantly more accurate analysis and prediction results using spatially complete data than when only limited observations are used. The mean ratios of the ensemble spread to the RMSE of the forecast for FullObs, MixObs, and PartObs in this baseline assimilation group were 0.85, 0.7, and 0.65, respectively. In this study, we used different resolutions for the forecast model (T63) and truth model (T85). The model error due to resolution difference was relatively large, leading to a relatively large RMSE of the forecast and a lower spread-to-RMSE ratio. We conducted identical resolution simulations and added the results to the appendix. For those simulations using T63 resolution for both the true and forecast models, the average forecast spread-RMSE ratio is close to one for all experimental groups.

2) SENSITIVITY TO OBSERVATION DENSITY

Figure 5 compares an instance of the streamfunction reconstruction based on the smaller amounts of observations in the Sparse and the Very-Sparse group. The CAE can reconstruct the streamfunction pattern shown in Figs. 5b and 5e, which resembles the FullObs in Fig. 2a very well, with very limited observations in Figs. 5a and 5d, though the intensity of the reconstructed field from the Very-Sparse Observation exhibits more errors than the Sparse Observation. The mixed

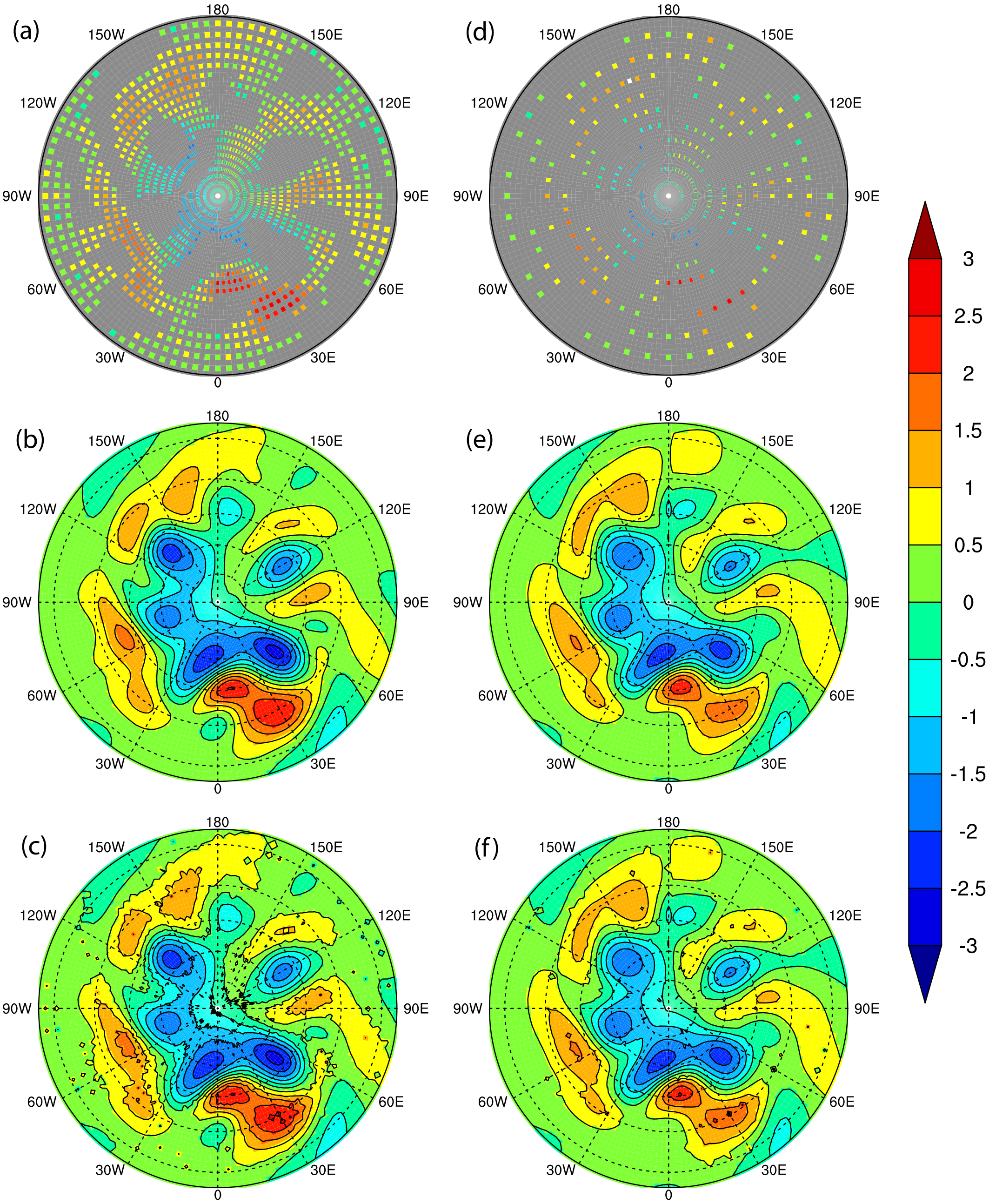


FIG. 5. The streamfunction (ψ) on a randomly selected day from (left) the sparse group and (right) the very sparse group; (a),(d) the observation of PartObsS and PartObsVS; (b),(e) the CAE reconstruct fields based on (a) and (d), respectively; (c),(f) the mixed observation of MixObsSC and MixObsVSC, respectively.

streamfunction based on 1/4 of the Dense Observation information (Fig. 5c) contains more details in the unobservable region and more random noise in the observable region compared with that based on the very sparse observation network, which has 1/16 of the observations in the Dense experiment (Fig. 5f).

To assess the impact of observation density on DA, the mean RMSEs of the analysis ensemble mean of the three experimental groups for the last 100 assimilation cycles are shown in Fig. 6. When we compare the three groups, we can see that the Dense group outperforms the other two sparse groups, which have less observation data. Analysis error increases as the observation density decreases, especially in experiments with spatially incomplete observations. PartObsS analysis error increases by 49% when compared to PartObs, and PartObsVS analysis error increases by more than twice as much. In the Dense group, the mean posterior RMSE shows the same result as Fig. 4d. After reducing the observation

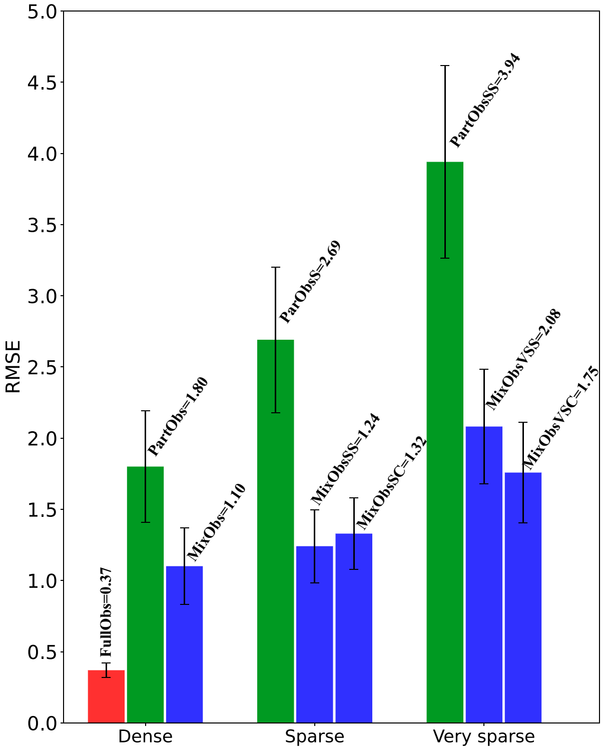


FIG. 6. The mean RMSE (10^6 m^2 s^{-1}) of the last-100-day posterior ensemble mean streamfunction for the Dense, Sparse, and Very-Sparse groups. The error bars indicate plus/minus one standard deviation.

density by a factor of 1/4, the analysis error of MixObsSS in the Sparse group, which is reconstructed from PartObsS, increases by only 13% when compared to MixObs analysis, which is much smaller than the PartObsS analysis. Though in the Very-Sparse group we only have 1/16 data of the dense observation, the streamfunction reconstructed by the CAE model in MixObsVSS still introduces notable improvement and lowers the analysis error by almost 50% compared with PartObsVS.

Also shown in Fig. 6 is the performance of assimilating the CAE reconstructed observations with full T63 resolution (MixObsSC and MixObsVSC). The reconstruction is based on coarsened 1/4 and 1/16 observations from PartObs, and assimilation experiments using the complete reconstructed observation yield different results than those using coarsened reconstruction data. The MixObsSC analysis exhibits a slightly larger error than the MixObsSS, probably due to the error introduced by the CAE reconstruction itself.

However, in the Very-Sparse group, the assimilation in MixObsVSC significantly lowers the analysis error compared with MixObsVSS. This implies that when we have a very limited amount of observation data, the benefit of deep learning reconstruction is more beneficial, even though it contains some inherent error.

Overall, even though mixed observation data perform differently in different groups, the benefits of deep learning augmentation are significant when compared to assimilating the spatially incomplete observations (1/16) only. Very interestingly, although MixObsVSC reconstructs from a very limited amount of observation, its analysis shows a slightly smaller error than the PartObs analysis in the Dense group. This further suggests the necessity of using deep learning to reconstruct pseudo-observations at locations without observations.

3) Sensitivity to ensemble size

Last, we examine the sensitivity of assimilation to ensemble size by comparing the result from the 80-member ensemble to that of a 10-member ensemble. The localization and inflation parameters tuned for 80-member ensembles are used for 10-member ensembles. The performance of 10-member ensemble could be improved with its tuned localization and inflation parameters. Figure 7 represents the mean analysis RMSE of assimilation experiments. It is expected that the analysis errors with 80 ensemble members are lower for the same observation dataset. However, the experiments that assimilate spatially incomplete observations without deep learning augmentation are more sensitive to ensemble size than the experiments with deep learning augmentation. For example, in the Dense group (Fig. 7a), shrinking the ensemble size increases the RMSE of the PartObs analysis by 62%; in contrast, the RMSE of the MixObs analysis only increases by 25%. The relatively larger increase in the FullObs analysis is most likely due to the FullObs' very low RMSE with the large ensemble, which is a "perfect" setting. Figure 7d depicts the analysis error from day 250 to day 350 in the Dense group experiments. The two PartObs experiments consistently exhibit a larger gap than the FullObs and MixObs groups.

In Fig. 7a, we can also find that even with only 10 ensemble members, the MixObs observation with full coverage can improve the assimilation performance compared with the PartObs analysis with the large ensemble size of 80. Figure 7b shows similar results. Assimilation in either MixObsSS or MixObsSC with only 10 ensemble members can reduce analysis errors to a level lower than that of the PartObsS analysis with 80 members. Furthermore, the MixObsSC analysis with 10 members yields a lower analysis error than the MixObsSS analysis, which contrasts with the results obtained with 80 members. This suggests that with limited ensemble members, reconstructed dense observations can obtain better analysis results than sparse observations even with the reconstruction errors.

In Fig. 7c for the Very-Sparse group, we can see that reducing the ensemble size has limited influence on the analysis result of MixObsVSC; the RMSE increases by only 15% which is much smaller than the RMSE increase of the MixObsVSS experiments. This demonstrates that the CAE reconstruction can obtain an analysis field of high accuracy even with very sparse observations.

Another interesting contrast shown in Figs. 7b and 7c is that when observation is spatially incomplete and sparse, using CAE reconstruction with the Dense resolution (MixObsSC and MixObsVSC) makes the analysis results significantly less sensitive to ensemble size than using reconstruction on the (very) sparse grid only. These findings imply that when observation is sparse, reconstructing spatially complete and more dense pseudo-observation via the CAE model may assist a

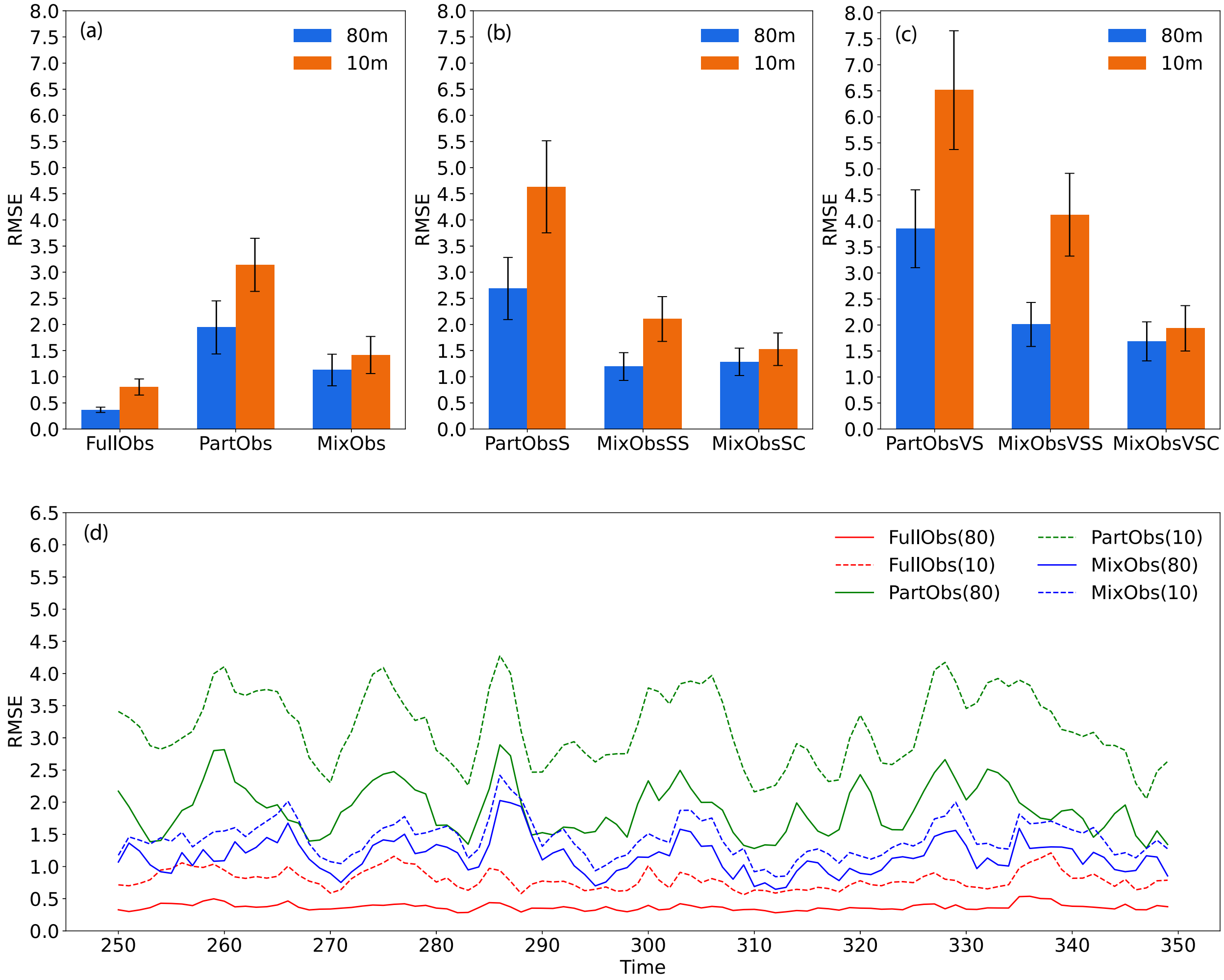


FIG. 7. The mean RMSE (10^6 m^2 s^{-1}) of the last-100-day posterior analysis streamfunction fields for the (a) Dense, (b) Sparse, and (c) Very-Sparse groups with 80 (blue) or 10 ensemble members (orange). (d) The RMSE time series of the last 100 days for the experiments in the Dense group with different ensemble sizes. Error bars in (a)–(c) indicate plus/minus one standard deviation.

small ensemble in achieving satisfactory performance, which may even be comparable to large ensemble assimilation with spatially incomplete but more dense observation. Thus, deep learning augmentation can potentially help reduce the computational cost of DA substantially.

5. Summary and discussion

The targeted scenario of this investigation is the assimilation of spatially incomplete observations, such as remote sensing data. In the case of radar observation, no information is available unless hydrometeors are present. Under clouds and precipitation, information on satellite infrared radiance is unavailable. The unobservable regions in such scenarios contain important environmental information and, in some cases, act as barriers to further DA improvement (Fabry and Meunier 2020).

Is it possible to "retrieve" the full physical fields from spatially incomplete remote sensing observations to enhance DA? Based on our idealized experiments, in which a deep neural network was trained to reconstruct the full physical fields from limited, incomplete observations, the answer appears promising. In real-world NWP situations, ensemble prediction data archives contain valuable information about the nonlinear relationship between model states and corresponding incomplete observations, the latter of which can be obtained using a (forward) observation operator. The deep learning model can be trained to approximate the inverse of the observation operator, which can be termed as the "reconstruction" operator.

In our experiments, we intentionally ran the benchmark simulation at a resolution higher than the forecast model's resolution. This configuration simulates the reality that our forecast model is not perfect, and thus the deep learning training dataset may be flawed. However, we successfully trained the deep learning model using a long-term "historical" simulation with the coarse resolution forecast model, which reconstructs full-coverage observations with high accuracy based on incomplete observations from the (high-resolution) benchmark simulation. Such reconstruction is still very skillful even when the input observation density is coarsened but full-resolution reconstruction is intended.

When observation data has the same resolution (T63) as the forecast model, the analysis streamfunction generated by assimilating the combination of partial observation (PartObs) and deep learning reconstruction (MixObs) is more accurate than the result from assimilating PartObs alone. The deep learning reconstruction also improves the ensembles' 1-day forecast error. When using the full-resolution reconstruction (MixObsSC and MixObsVSC), the benefits of using mixed observations do not degrade significantly when observation data becomes very sparse; in contrast, the analysis generated by assimilating partially available observations is more sensitive to observation density. The advantage of deep learning augmentation even extends to small ensemble assimilation, therefore it has a promising potential in reducing the computational cost of ensemble DA.

The accuracy of real case applications of DDA will be bottlenecked not only by the capacity of the deep neural networks but also by the reliability of our NWP models, which provide the training datasets. The convolutional autoencoder used in this study can be generalized to three-dimensional datasets by using three-dimensional convolutional and transposed convolutional layers. Multiple variables can be regarded as multiple network channels. Though our application is based on the barotropic vorticity model which has relatively limited subspace dimentionality,[2] and is therefore not extremely difficult for deep learning reconstruction, for practical applications, the specific structure of the deep learning model can be flexible, and the continuing advancement of deep learning will certainly provide us with more powerful tools (e.g., Jam et al. 2021; Kang et al. 2021).

The reliability of operational NWP models has been improving over the decades along with steady advances in data assimilation systems (Blayo et al. 2014; Geer et al. 2018; Pu and Kalnay 2019). As a result, the quality of training data derived from historical NWP forecasts and reanalyses is expected to be adequate for training deep learning models. Therefore, the practical barrier to realizing DDA's full potential may be the large quantity, rather than the quality, of geophysical datasets because of the expected demanding computational cost. It is common to use more than ~10 000 samples in training deep learning models (e.g., Weyn et al. 2020; Ham et al. 2021; Pathak et al. 2022). Including high-resolution three-dimensional atmospheric state data for such many sample states would make the computational cost infeasible. The advancement of distributed deep learning (Dean et al. 2012; Mayer and Jacobsen 2020) could provide a pathway to operational applications. However, for research and preliminary testing, it might be desirable to first focus on a limited region, instead of the global atmosphere. Another strategy to downsize training datasets might be coarsening the original datasets to lower horizontal and vertical resolutions. For example, in the development of FourCastNet, Pathak et al. (2022) used the European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis version 5 (ERA5; C3S 2017) data from 1979 to 2015 for model training, but to make the computational cost affordable, they only used 6-hourly data on five vertical levels, instead of the full ERA5 dataset with hourly intervals and on 37 levels. However, because the original datasets are generated with a high-resolution model, nonlinear relationships between observation and model states are still preserved and consistent.

Though not shown in this paper, the same CAE structure used here has been tested and shown to be capable of mapping a spatially incomplete vorticity field to its corresponding full streamfunction field with accuracy comparable to the streamfunction-mapping CAE we discussed. Because remote sensing data cannot be reconstructed in their unobservable regions, such mapping between variables will be required for more realistic investigations. In the future, we will investigate the application of deep learning augmentation to more realistic DA cases.

Acknowledgments. The authors thank Drs. Gad Levy, Pavel Sakov, Jeffrey Anderson, and two anonymous reviewers for their constructive comments. Y.W. and X.S. are supported by the Research Grants Council of Hong Kong SAR, China (Projects AoE/E-603/18 and HKUST 16301721). J.F. is supported by the Science and Technology Planning Project of Guangdong Province, China (2018A050501004). L.L. is supported by the National Outstanding Youth Science Fund Project of National Natural Science Foundation of China (Grant 41922036).

Data availability statement. The original barotropic model code was provided by the GFDL and downloaded from https://www.gfdl.noaa.gov/idealized-spectral-models-quickstart/. The TensorFlow code for training the CAE can be found at https://github.com/shixm-cloud/DA2.

APPENDIX

Identical Resolution Model Experiments

Here we document some experiment results when the truth and forecast model resolutions are identical. Figure A1 shows the forecast and analysis RMSE of the ensemble means of the three assimilation experiments (FullObs, PartObs, and MixObs) against the truth data in the case of using identical truth and forecast model resolution of 96 × 192 (T63). This shows that FullObs outperforms MixObs and the PartObs in both the analysis and forecast results. The MixObs analysis error is comparable with the PartObs. However, MixObs forecast RMSE is still slightly better than the PartObs in most cases. We can conclude that even though the analysis results are not improved substantially, we can get notably more accurate forecast results with complete observations from the CAE reconstruction.

Figure A2 compares the analysis and forecast results of the ensemble means of the Very-Sparse Observations (PartObsVS), which has coarsened 1/16 of data from the PartObs and two

[2] Empirical orthogonal function (EOF) analysis suggests that the first 30 EOF modes can explain 78% of the variance in the streamfunction field, and the first 100 EOF modes explain 98% of total variance.

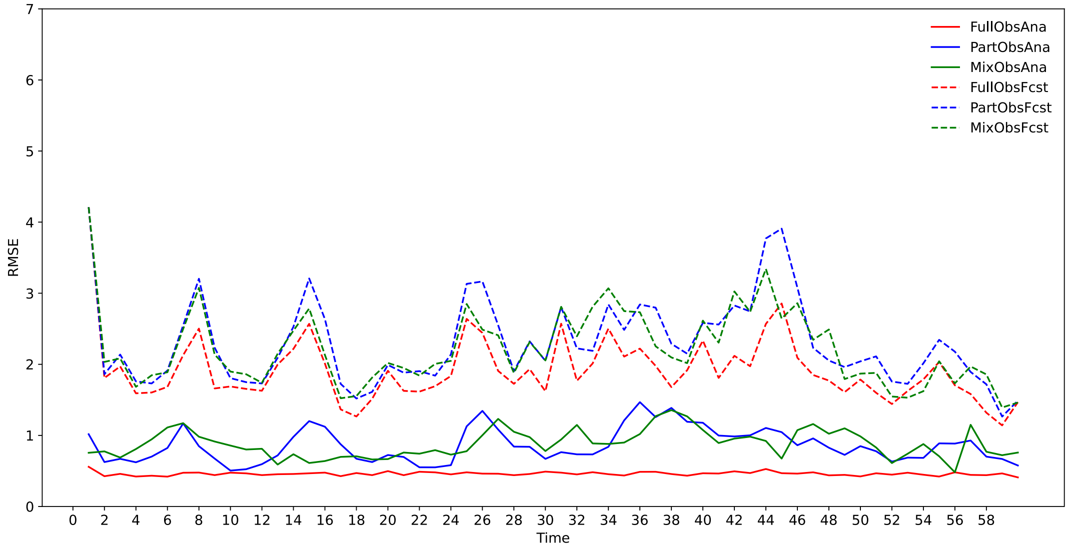


FIG. A1. The root-mean-square error (RMSE) (10^6 m^2 s^{-1}) of the prior ensemble (forecast) mean (dashed lines) and posterior ensemble (analysis) mean (solid lines) during the assimilation cycles (60 days) for the FullObs (red lines), PartObs (blue lines), and the MixObs (green lines) experiments. A total of 80 ensemble members are used in each experiment.

mixed observations (MixObsVSC and MixObsVSS) reconstructed based on the PartObsVS with the same true and forecast model resolution (T63). The two CAE-augmented mixed observations exhibit lower analysis RMSE than the PartObs analysis during almost all the cycling periods. It seems that the MixObsVSC shows slightly lower analysis RMSE than MixObsVSS which is similar to the result of the sparse group in Fig. 6. When we focus on the forecast RMSE of the three kinds of observations, the mixed observations showed lower prior RMSE than the PartObs, same as in Fig. 4d.

Therefore, we demonstrate that in the case of using an identical truth and forecast model resolution, the assimilation can still benefit from the CAE reconstruction, and the advantages are substantial in the case of having very sparse observations.

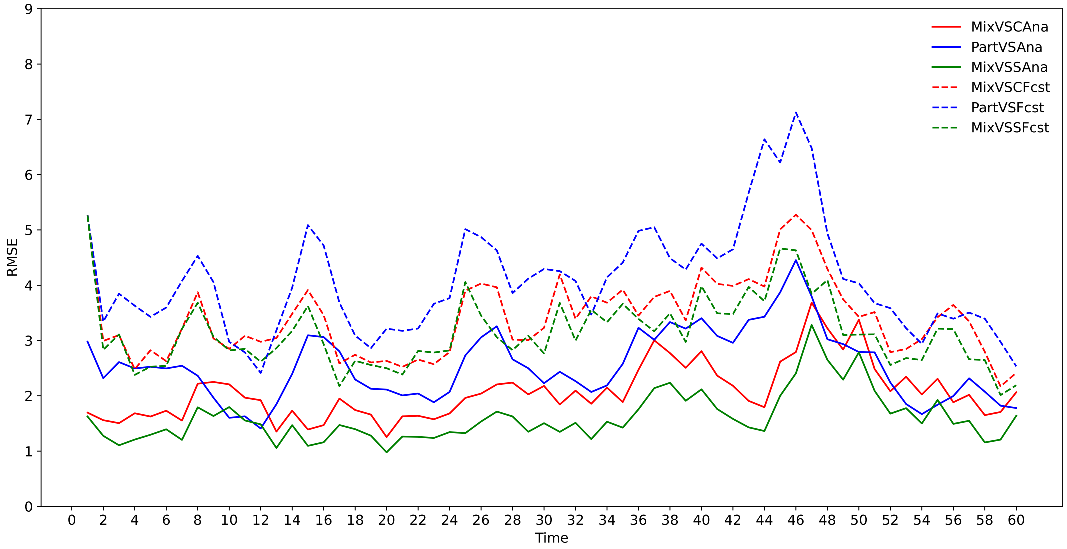


FIG. A2. The root-mean-square error (RMSE) (10^6 m^2 s^{-1}) of the prior ensemble (forecast) mean (dashed lines) and posterior ensemble (analysis) mean (solid lines) during the assimilation cycles (60 days) for the PartObsVS (blue lines), MixObsVSC (red lines), and the MixObsVSS (green lines) experiments. A total of 80 ensemble members are used in each experiment.

REFERENCES

Aires, F., W. B. Rossow, N. A. Scott, and A. Chédin, 2002: Remote sensing from the infrared atmospheric sounding interferometer instrument 2. Simultaneous retrieval of temperature, water vapor, and ozone atmospheric profiles. *J. Geophys. Res.*, **107**, 4620, https://doi.org/10.1029/2001JD001591.

Alley, R. B., K. A. Emanuel, and F. Zhang, 2019: Advances in weather prediction. *Science*, **363**, 342–344, https://doi.org/10.1126/science.aav7274.

Anderson, J. L., and S. L. Anderson, 1999: A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Mon. Wea. Rev.*, **127**, 2741–2758, https://doi.org/10.1175/1520-0493(1999)127<2741:AMCIOT>2.0.CO;2.

Blayo, É., M. Bocquet, E. Cosme, and L. F. Cugliandolo, 2014: Advanced data assimilation for geosciences: Lecture notes of the Les Houches School of Physics: Special Issue, June 2012. Oxford University Press, https://doi.org/10.1093/acprof:oso/9780198723844.001.0001.

Blyverket, J., P. D. Hamer, L. Bertino, C. Albergel, D. Fairbairn, and W. A. Lahoz, 2019: An Evaluation of the EnKF vs. EnOI and the assimilation of SMAP, SMOS and ESA CCI Soil moisture data over the contiguous U.S. *Remote Sens.*, **11**, 478, https://doi.org/10.3390/rs11050478.

Bobylev, L. P., E. V. Zabolotskikh, L. M. Mitnik, and M. L. Mitnik, 2009: Atmospheric water vapor and cloud liquid water retrieval over the Arctic Ocean using satellite passive microwave sensing. *IEEE Trans. Geosci. Remote Sens.*, **48**, 283–294, https://doi.org/10.1109/TGRS.2009.2028018.

Burgers, G., P. Jan van Leeuwen, and G. Evensen, 1998: Analysis scheme in the ensemble Kalman filter. *Mon. Wea. Rev.*, **126**, 1719–1724, https://doi.org/10.1175/1520-0493(1998)126<1719:ASITEK>2.0.CO;2.

Caron, J.-F., and M. Buehner, 2018: Scale-dependent background error covariance localization: Evaluation in a global deterministic weather forecasting system. *Mon. Wea. Rev.*, **146**, 1367–1381, https://doi.org/10.1175/MWR-D-17-0369.1.

Casas, C. Q., R. Arcucci, P. Wu, C. Pain, and Y.-K. Guo, 2020: A reduced order deep data assimilation model. *Physica D*, **412**, 132615, https://doi.org/10.1016/j.physd.2020.132615.

C3S, 2017: ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate. Copernicus Climate Change Service Climate Data Store (CDS), accessed 13 March 2021, https://cds.climate.copernicus.eu/.

Dean, J., and Coauthors, 2012: Large scale distributed deep networks. *Advances in Neural Information Processing Systems 25: 26th Annual Conf. on Neural Information Processing Systems 2012 (NIPS 2012)*, Lake Tahoe, NV, Neural Information Processing Systems Foundation, Inc. (NIPS), 1223–1231, https://papers.nips.cc/paper/2012/hash/6aca97005c68f1206823815f66102863-Abstract.html.

Errico, R. M., P. Bauer, and J.-F. Mahfouf, 2007: Issues regarding the assimilation of cloud and precipitation data. *J. Atmos. Sci.*, **64**, 3785–3798, https://doi.org/10.1175/2006JAS2044.1.

Fabry, F., and V. Meunier, 2020: Why are radar data so difficult to assimilate skillfully? *Mon. Wea. Rev.*, **148**, 2819–2836, https://doi.org/10.1175/MWR-D-19-0374.1.

Gaspari, G., and S. E. Cohn, 1999: Construction of correlation functions in two and three dimensions. *Quart. J. Roy. Meteor. Soc.*, **125**, 723–757, https://doi.org/10.1002/qj.49712555417.

Geer, A. J., and Coauthors, 2018: All-sky satellite data assimilation at operational weather forecasting centres. *Quart. J. Roy. Meteor. Soc.*, **144**, 1191–1217, https://doi.org/10.1002/qj.3202.

Ham, Y. G., J. H. Kim, E. S. Kim, and K. W. On, 2021: Unified deep learning model for El Niño/Southern Oscillation forecasts by incorporating seasonality in climate data. *Sci. Bull.*, **66**, 1358–1366, https://doi.org/10.1016/j.scib.2021.03.009.

Jam, J., C. Kendrick, K. Walker, V. Drouard, J. G. S. Hsu, and M. H. Yap, 2021: A comprehensive review of past and present image inpainting methods. *Comput. Vis. Image Underst.*, **203**, 103147, https://doi.org/10.1016/j.cviu.2020.103147.

Jin, J., H. X. Lin, A. Segers, Y. Xie, and A. Heemink, 2019: Machine learning for observation bias correction with application to dust storm data assimilation. *Atmos. Chem. Phys.*, **19**, 10 009–10 026, https://doi.org/10.5194/acp-19-10009-2019.

Jung, Y., G. Zhang, and M. Xue, 2008: Assimilation of simulated polarimetric radar data for a convective storm using the ensemble Kalman filter. Part I: Observation operators for reflectivity and polarimetric variables. *Mon. Wea. Rev.*, **136**, 2228–2245, https://doi.org/10.1175/2007MWR2083.1.

Kang, S. K., S. A. Shin, S. Seo, M. S. Byun, D. Y. Lee, Y. K. Kim, D. S. Lee, and J. S. Lee, 2021: Deep learning-based 3D inpainting of brain MR images. *Sci. Rep.*, **11**, 1673, https://doi.org/10.1038/s41598-020-80930-w.

Mao, X.-J., C. Shen, and Y.-B. Yang, 2016: Image restoration using convolutional auto-encoders with symmetric skip connections. arXiv, 1606.08921, https://arxiv.org/abs/1606.08921.

Mayer, R., and H.-A. Jacobsen, 2020: Scalable deep learning on distributed infrastructures: Challenges, techniques, and tools. *ACM Comput. Surv.*, **53**, 1–37, https://doi.org/10.1145/3363554.

Miyoshi, T., and K. Kondo, 2013: A multi-scale localization approach to an ensemble Kalman filter. *SOLA*, **9**, 170–173, https://doi.org/10.2151/sola.2013-038.

Pathak J., and Coauthors, 2022: FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators. arXiv, 2202.11214, https://arxiv.org/abs/2202.11214.

Pu, Z., and E. Kalnay, 2019: Numerical weather prediction basics: Models, numerical methods, and data assimilation. *Handbook of Hydrometeorological Ensemble Forecasting*, Q. Duan, Eds., Springer, 67–97.

Sakov, P., and P. R. Oke, 2008: Implications of the form of the ensemble transformation in the ensemble square root filters. *Mon. Wea. Rev.*, **136**, 1042–1053, https://doi.org/10.1175/2007MWR2021.1.

Simmons, A. J., and A. Hollingsworth, 2002: Some aspects of the improvement in skill of numerical weather prediction. *Quart. J. Roy. Meteor. Soc.*, **128**, 647–677, https://doi.org/10.1256/003590002321042135.

Sodhi, J. S., and F. Fabry, 2020: Multiscale assimilation of radar reflectivity. *24th Conf. on Integrated Observing and Assimilation Systems for the Atmosphere, Oceans, and Land Surface (IOAS-AOLS)*, Boston, MA, Amer. Meteor. Soc., 4A.4, https://ams.confex.com/ams/2020Annual/meetingapp.cgi/Paper/363970.

Tippett, M. K., J. L. Anderson, C. H. Bishop, T. M. Hamill, and J. S. Whitaker, 2003: Ensemble square root filters. *Mon. Wea. Rev.*, **131**, 1485–1490, https://doi.org/10.1175/1520-0493(2003)131<1485:ESRF>2.0.CO;2.

Vallis, G. K., E. P. Gerber, P. J. Kushner, and B. A. Cash, 2004: A mechanism and simple dynamical model of the North Atlantic Oscillation and annular modes. *J. Atmos. Sci.*, **61**, 264–280, https://doi.org/10.1175/1520-0469(2004)061<0264:AMASDM>2.0.CO;2.

Wang, Z., E. P. Simoncelli, and A. C. Bovik, 2003: Multiscale structural similarity for image quality assessment. *The Thrity-Seventh Asilomar Conf. on Signals, Systems & Computers*, Institute of Electrical and Electronics Engineers, Vol. 2, 1398–1402.

Weyn, J. A., D. R. Durran, and R. Caruana, 2020: Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *J. Adv. Model. Earth Syst.*, **12**, e2020MS002109, https://doi.org/10.1029/2020MS002109.

Whitaker, J. S., and T. M. Hamill, 2002: Ensemble data assimilation without perturbed observations. *Mon. Wea. Rev.*, **130**, 1913–1924, https://doi.org/10.1175/1520-0493(2002)130<1913:EDAWPO>2.0.CO;2.

Xie, J., L. Xu, and E. Chen, 2012: Image denoising and inpainting with deep neural networks. *Adv. Neural Info. Process. Syst.*, **25**, 341–349.

Ying, Y., 2020: Assimilating observations with spatially correlated errors using a serial ensemble filter with a multiscale approach. *Mon. Wea. Rev.*, **148**, 3397–3412, https://doi.org/10.1175/MWR-D-19-0387.1.

Zheng, M., and Coauthors, 2021: Data gaps within atmospheric rivers over the northeastern Pacific. *Bull. Amer. Meteor. Soc.*, **102**, E492–E524, https://doi.org/10.1175/BAMS-D-19-0287.1.