



Ensemble data assimilation to diagnose AI-based weather prediction models: a case with ClimaX version 0.3.1

Shunji Kotsuki^{1,2,3}, Kenta Shiraishi⁴, and Atsushi Okazaki^{1,2}

¹Institute for Advanced Academic Research, Chiba University, Chiba, Japan

²Center for Environmental Remote Sensing, Chiba University, Chiba, Japan

³Research Institute of Disaster Medicine, Chiba University, Chiba, Japan

⁴Graduate School of Science and Engineering, Chiba University, Chiba, Japan

Correspondence: Shunji Kotsuki (shunji.kotsuki@chiba-u.jp)

Received: 4 October 2024 – Discussion started: 15 October 2024

Revised: 23 May 2025 – Accepted: 3 July 2025 – Published: 15 October 2025

Abstract. Artificial intelligence (AI)-based weather prediction research is growing rapidly and has shown to be competitive with advanced dynamic numerical weather prediction (NWP) models. However, research combining AI-based weather prediction models with data assimilation remains limited, partially because long-term sequential data assimilation cycles are required to evaluate data assimilation systems. This study proposes using ensemble data assimilation for diagnosing AI-based weather prediction models and marked the first successful implementation of the ensemble Kalman filter with AI-based weather prediction models. Our experiments with an AI-based model, ClimaX, demonstrated that the ensemble data assimilation cycled stably for the AI-based weather prediction model using covariance inflation and localization techniques within the ensemble Kalman filter. While ClimaX showed some limitations in capturing flow-dependent error covariance compared to dynamical models, the AI-based ensemble forecasts provided reasonable and beneficial error covariance in sparsely observed regions. In addition, ensemble data assimilation revealed that error growth based on ensemble ClimaX predictions was weaker than that of dynamical NWP models, leading to higher inflation factors. A series of experiments demonstrated that ensemble data assimilation can be used to diagnose properties of AI weather prediction models, such as physical consistency and accurate error growth representation.

1 Introduction

The intensification of weather-induced disasters due to climate change is becoming increasingly severe worldwide (e.g., Jonkman et al., 2024). In a recent risk report, the World Economic Forum (2023) indicated that extreme weather is among the most severe global threats. To address extreme-weather events such as torrential heavy rains and heat waves, further advancements in weather forecasting are essential. There are two essential components for accurate weather forecasting: (1) numerical weather prediction (NWP) models that forecast future weather based on initial conditions and (2) data assimilation, which integrates atmospheric observation data to estimate initial conditions for subsequent forecasts by NWP models.

Since NVIDIA issued the first artificial intelligence (AI) weather prediction model, FourCastNet, which was competitive with dynamical NWP models, in February 2022 (Pathak et al., 2022; Bonev et al., 2023), deep-learning-based weather prediction research has shown rapid growth. A number of AI weather prediction models have been proposed, mainly by private information and technology (IT) companies, such as GraphCast by Google DeepMind (Lam et al., 2023), Pangu-Weather by Huawei (Bi et al., 2023), ClimaX and Stormer by Microsoft (Nguyen et al., 2023), and Aurora by Microsoft (Bodnar et al., 2025). These machine learning approaches have been shown to be competitive with state-of-the-art NWP models (e.g., Kochkov et al., 2024). Progress in AI-based weather prediction has been supported by the expansion of benchmark data and evaluation algorithms, such as WeatherBench (Rasp et al., 2020, 2024). Notably, most AI-

based weather prediction models, including Pang-Weather, ClimaX, Stormer, and FourCastNet, use the Vision Transformer (ViT) neural network architecture (Vaswani et al., 2017; Dosovitskiy et al., 2021). The ViT, which has been explored in language models and image classifications, has been demonstrated to be effective in weather prediction as well.

However, research that couples AI-based weather prediction models with data assimilation remains limited. This limitation is partially due to the fact that long-term sequential data assimilation experiments are needed for the evaluation of data assimilation systems, in contrast to weather prediction tasks that allow for parallel learning using benchmark data. Conventional data assimilation methods used in NWP systems can be categorized into three groups: variational methods, ensemble Kalman filters, and particle filters. There are strong mathematical similarities between neural networks and variational data assimilation, both of which minimize their cost functions using their differentiable models. Because auto-differentiation codes are always available for neural-network-based AI models, AI weather prediction models are considered compatible with variational data assimilation methods, as in Xiao et al. (2024) and Adrian et al. (2025). On the other hand, recent studies have started to solve the inverse problem inherent in data assimilation using deep neural networks (McCabe and Brown, 2021; Chen et al., 2024; Bocquet et al., 2024; Luk et al., 2024; Vaughan et al., 2024). Some studies have employed ensemble Kalman filters for data-driven models (Hamilton et al., 2016; Penny et al., 2022; Chattopadhyay et al., 2022, 2023). However, no study has succeeded in employing ensemble Kalman filtering with global AI models of the atmosphere. Since AI models require significantly lower computational costs compared to dynamical NWP models, AI models offer benefits for ensemble-based methods, such as ensemble Kalman filters (EnKFs) and particle filters. Ensemble data assimilation at the global scale also allows us to assess the capability of data assimilation with AI models to handle spatially inhomogeneous observation networks and to maintain physically consistent multivariate error covariance across the entire atmosphere.

This study proposes using ensemble data assimilation for diagnosing AI-based weather prediction models. For that purpose, this study marks the first successful implementation of ensemble Kalman filter experiments with an AI weather prediction model to the best of the authors' knowledge. We applied the ViT-based ClimaX (Nguyen et al., 2023) to data assimilation experiments using the available source code and experimental environments with necessary modifications. For data assimilation, we applied the local ensemble transform Kalman filter (LETKF) (Hunt et al., 2007), which is among the most widely used data assimilation methods in operational NWP centers, such as the European Centre for Medium-Range Weather Forecasts (ECMWF), the Deutscher Wetterdienst (DWD), and the Japan Meteorologi-

cal Agency (JMA). Using the coupled ClimaX–LETKF data assimilation system, we investigated several key aspects of AI-based weather prediction models, including whether the data assimilation cycles stably for the ClimaX AI weather prediction model using ensemble Kalman filters and whether AI-based ensemble weather prediction accurately represents flow-dependent background error variance and covariance. We also investigated whether techniques such as covariance inflation and localization, which are conventionally used in EnKFs for dynamical NWP models, are effective for AI weather prediction models. By addressing these research questions, we aim to advance the integration of AI weather prediction models with data assimilation techniques toward the development of more accurate weather forecasting. While this study primarily aims to use ensemble data assimilation for diagnosing AI-based weather prediction models, our research also represents an important step toward enabling real-time updates of AI weather models with meteorological observations.

The rest of the paper is organized as follows. Sect. 2 describes the methods and experiments, and Sect. 3 presents the results. Finally, Sect. 4 provides a discussion and summary.

2 Methods and experiments

2.1 ClimaX model

ClimaX (Nguyen et al., 2023) is a ViT-based AI weather prediction model for the global atmosphere. Variable tokenization and variable aggregation are key components of the ClimaX architecture on ViT, as they provide flexibility and generality. This study used the low-resolution version of ClimaX (version 0.3.1), with 64 and 32 zonal and meridional grid points, respectively, corresponding to a spatial resolution of $5.625^\circ \times 5.625^\circ$. The vertical model level was set at seven (900, 850, 700, 600, 500, 250, and 50 hPa).

By default, ClimaX is set to be trained against only five variables: geopotential at 500 hPa, temperature at 850 hPa, temperature at 2 m, zonal wind at 10 m, and meridional wind at 10 m. We updated ClimaX for data assimilation, which allowed the AI model to produce variables required for subsequent forecasts (Table 1). The updated ClimaX has state vectors, including zonal wind, meridional wind, temperature, specific humidity, and geopotential at seven vertical layers along with three surface variables: 10 m zonal wind, 10 m meridional wind, and 2 m temperature. We also diagnosed surface pressure, which is a required input for data assimilation, based on geopotential and surface elevation. Figure 1 shows the training curves of the default and updated ClimaX models verified against WeatherBench data (Rasp et al., 2020). Data for the period 2006–2015 were used for training, and data for 2016 were used for validation. Here we re-trained ClimaX entirely with the additional outputs (i.e., no transfer learning). It took approximately 4 h with four GPUs

of NVIDIA RTX 6000Ada. Anomaly correlation coefficients increased, and root mean square errors (RMSEs) decreased, as shown in Fig. 1, indicating successful training of the updated ClimaX model. Because more variables were predicted by the updated ClimaX than by the default ClimaX, more training steps were required.

2.2 Local ensemble transform Kalman filter (LETKF)

The LETKF is among the most widely used data assimilation methods in operational NWP centers, such as ECMWF, DWD, and JMA. The LETKF simultaneously computes analysis equations at every model grid point with the assimilation of surrounding observations within the localization cut-off radius. The ClimaX–LETKF system was developed based on the SPEEDY–LETKF system (Kotsuki et al., 2022) by replacing the SPEEDY weather prediction model with ClimaX. Our future research can readily be expanded to particle filter experiments because the Kotsuki et al. (2022) system includes local particle filters in addition to the LETKF.

Let $\mathbf{X}_t \equiv \{\mathbf{x}_t^{(1)}, \dots, \mathbf{x}_t^{(m)}\}$ be an ensemble state matrix, whose ensemble mean and perturbation are given by $\bar{\mathbf{x}}_t \in \mathbb{R}^n$ and $\delta\mathbf{X}_t \equiv \{\mathbf{x}_t^{(1)} - \bar{\mathbf{x}}_t, \dots, \mathbf{x}_t^{(m)} - \bar{\mathbf{x}}_t\} \in \mathbb{R}^{n \times m}$, respectively. Here, n and m are the system and ensemble sizes. The superscript (i) and subscript t denote the i th ensemble member and the time, respectively. The EnKF, including LETKF, estimates the error covariance $\mathbf{P} \in \mathbb{R}^{n \times n}$ according to sample estimates based on ensemble perturbation:

$$\mathbf{P} \approx \frac{1}{m-1} \delta\mathbf{X} \delta\mathbf{X}^T. \quad (1)$$

The analysis update equation of the LETKF is given by

$$\mathbf{X}_t^a = \bar{\mathbf{x}}_t^b \cdot \mathbf{1} + \delta\mathbf{X}_t^b \tilde{\mathbf{P}}_t^a \left(\mathbf{Y}_t^b\right)^T \mathbf{R}_t^{-1} \left(\mathbf{y}_t^o - \overline{H_t(\mathbf{X}_t^b)}\right) \cdot \mathbf{1} + \left[(m-1)\tilde{\mathbf{P}}_t^a\right]^{1/2}, \quad (2)$$

$$\tilde{\mathbf{P}}_t^a = \left[\frac{(m-1)}{\beta} \mathbf{I} + \left(\mathbf{Y}_t^b\right)^T \mathbf{R}_t^{-1} \mathbf{Y}_t^b \right]^{-1}, \quad (3)$$

where $\tilde{\mathbf{P}}$ is the error covariance matrix in the ensemble space ($\in \mathbb{R}^{m \times m}$), $\mathbf{Y} \equiv \mathbf{H} \delta\mathbf{X}$ is the ensemble perturbation matrix in the observation space ($\in \mathbb{R}^{p \times m}$), $\mathbf{R}$ is the observation error covariance matrix ($\in \mathbb{R}^{p \times p}$), $\mathbf{y}$ is the observation vector ($\in \mathbb{R}^p$), H is the observation operator that may be nonlinear, $\mathbf{H} \in \mathbb{R}^{p \times n}$ is the Jacobian of the linear observation operator matrix, and $\mathbf{1}$ is a row vector whose elements are all 1 ($\in \mathbb{R}^m$). Here, p is the number of observations. The superscripts o, b, and a denote the observation, background, and analysis, respectively. The scalar β is a multiplicative inflation factor which inflates the background error covariance such that $\mathbf{P}_t^b \rightarrow (1 + \beta)\mathbf{P}_t^b$. This study uses the Miyoshi (2011) approach, which estimates spatially varying inflation factors adaptively based on observation–space statistics (Desroziers et al., 2005).

Localization is a practically important technique for EnKFs to eliminate long-range erroneous correlations due to the sample estimates of $\mathbf{P}$ with a limited ensemble size (Houtekamer and Zhang, 2016). Although a larger localization can spread observation data information for grid points distant from observations, a larger localization scale can yield suboptimal error covariance because of sampling errors. The LETKF inflates the observation error variance to realize the localization (Hunt et al., 2007) whose function is given by

$$l = \begin{cases} \exp\left[-\frac{1}{2} \left\{ (d_h/L_h)^2 + (d_v/L_v)^2 \right\}\right] & \text{if } d_h < 2\sqrt{10/3}L_h \text{ and } d_v < 2\sqrt{10/3}L_v \\ 0 & \text{else} \end{cases}, \quad (4)$$

where l is the localization function, and its inverse l^{-1} is multiplied to inflate $\mathbf{R}$ for the localization. Horizontal and vertical distances (km and log(Pa)) from the analysis grid point to the observation are defined by d_h and d_v , where subscripts h and v denote horizontal and vertical, respectively. Here, L_h and L_v are tunable horizontal and vertical localization scales (km and log(Pa)). The vertical localization scale L_v was set at 1.0 (log Pa) following the method of Kotsuki et al. (2022). Sensitivity to the horizontal localization scale for $L_h = 400, 500, 600, 700$, and 800 km is investigated in subsequent experiments.

2.3 Data assimilation experiments

In this study, all experiments were conducted as simulation experiments by generating observation data from WeatherBench with additions of Gaussian random noises. Although the real observation data were not directly assimilated, the assimilated observations reflect the real atmosphere in this study, in contrast to observing system simulation experiments. To approximate real-world scenarios, we considered radiosonde-like observations to generate atmospheric observation profiles for observing stations (Fig. 2). At observing stations, temperature and zonal and meridional winds were observed at all seven layers, whereas specific humidity was observed at the first to fourth layers. Table 1 shows the standard deviations of the observation errors. The network of observing stations and observation error standard deviations were consistent with those of the SPEEDY–LETKF experiments (Kotsuki et al., 2022; Kotsuki and Bishop, 2022). Observation data were produced at 6 h intervals such that the data assimilation interval was also 6 h. Since the observation data were generated directly at the model grid points, the observation operator is a linear operator composed only of 0.0 and 1.0.

We employed a series of data assimilation experiments over the year 2017, which is not used for training and validation of ClimaX. The ensemble size is 20. Their initial conditions for 00:00 UTC, 1 January 2017, were taken from WeatherBench data in 2006, which were sampled every 12 h

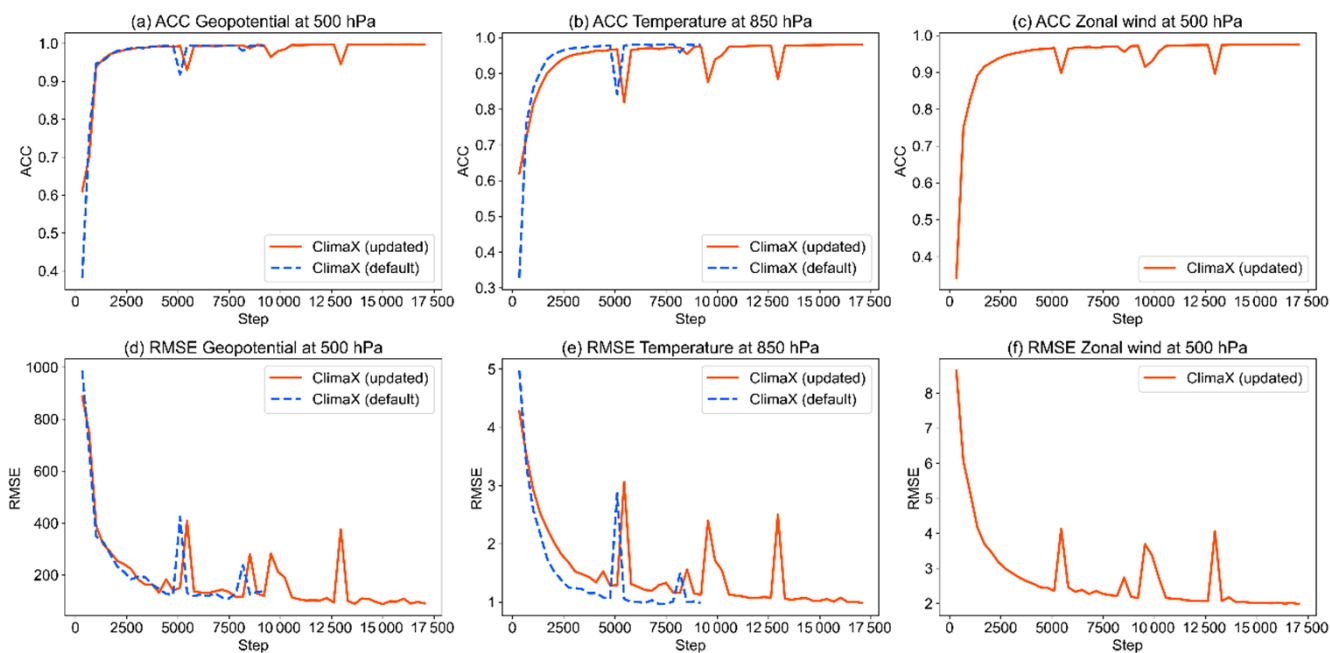


Figure 1. Training curves for the default and updated ClimaX models (dashed blue and solid orange lines) verified against WeatherBench data in 2016, as a function of the number of training steps. Each training step includes 64 training data in a mini batch. Panels (a)–(c) and (d)–(f) show anomaly correlation coefficients (ACCs) and root mean square errors (RMSEs). Panels (a) and (d), (b) and (e), and (c) and (f) are geopotential at 500 hPa ($\text{m}^2 \text{s}^{-2}$), temperature at 850 hPa, and zonal wind at 500 hPa. There are no dashed blue lines in panels (c) and (f) because the default ClimaX model does not predict zonal wind at 500 hPa.

Table 1. Variables of the ClimaX model used in this study. ClimaX requires input variables to predict output variables. Observation (Obs) variables are assimilated with the associated error standard deviation (error SD).

Symbol	Variable	Unit	Input	Output	Obs	Error SD	Height
U	Zonal wind	m s^{-1}	X	X	X	1.0	925, 850, 700, 600, 500, 250, 50 (hPa)
V	Meridional wind	m s^{-1}	X	X	X	1.0	
T	Temperature	K	X	X	X	1.0	
Q	Specific humidity	kg kg^{-1}	X	X	X*	0.1	
Geo	Geopotential	$\text{m}^2 \text{s}^{-2}$	X	X			
U10m	10 m zonal wind	m s^{-1}	X	X			10 m
V10m	10 m meridional wind	m s^{-1}	X	X			10 m
T2m	2 m temperature	m s^{-1}	X	X			2 m
Ps	Surface pressure	hPa			X	1.0	Surface
Elev	Surface elevation	m	X				Surface
Long	Longitude	degree	X				–
Lat	Latitude	degree	X				–
Mask	Land–sea mask	1 or 0	X				–

* Specific humidity is observed up to the fourth model level (i.e., 925, 850, 700, and 600 hPa).

from 00:00 UTC, 1 January 2006. Data assimilation experimental results were verified against WeatherBench data.

It should be noted that we were unable to conduct observation system simulation experiments (OSSEs), which require a natural run by ClimaX. This is because ClimaX could not produce long-term forecasts within our experimental config-

urations. A typical example is shown in Fig. 3. The forecasted temperature fields of ClimaX eventually began to deviate from the WeatherBench data with the continuation of 6 h forecasts. Ultimately, ClimaX produced meteorologically unrealistic weather fields, as demonstrated by the very low temperatures in the Pacific Ocean. Because AI models can-

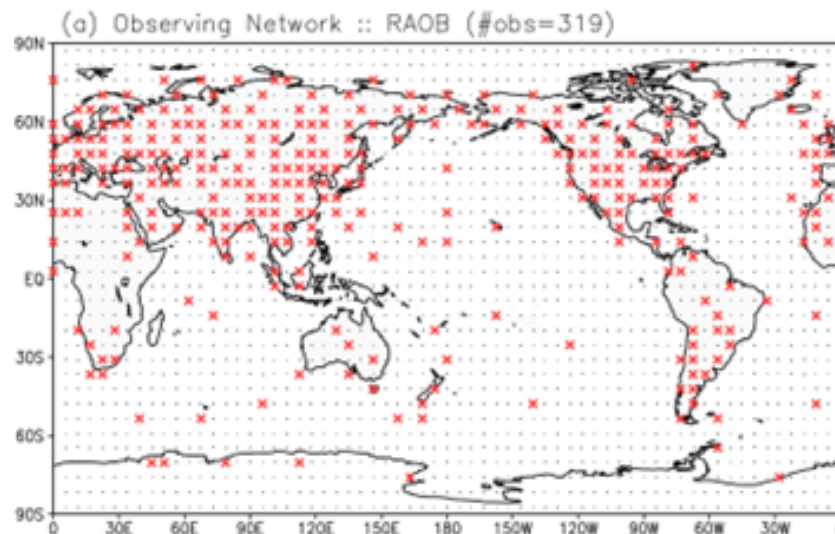


Figure 2. The observing network. The small black dots and red crosses represent model grid points and observing points, respectively.

not learn physical laws in the absence of specific treatments, they are more likely to produce unrealistic weather fields under previously unencountered weather conditions. In other words, this suggests that ClimaX is unable to return to a meteorologically plausible attractor (or trajectory), while data assimilation enables ClimaX to synchronize with the real atmosphere, as shown in Sect. 3. This property in AI models was theoretically demonstrated by Adrian et al. (2025), who showed that long-term filter accuracy can be achieved with surrogate models if the models can provide accurate short-term forecasts. Applying neural networks that are informed or constrained by physical laws would be necessary to conduct observation system simulation experiments for AI-based weather prediction models.

3 Results

Figure 4 presents the time series of global mean root mean square errors (RMSEs) for temperature and geopotential height at the fifth model level, with four different horizontal localization scales (L_h). After the initiation of data assimilation, all experiments showed reductions in analysis errors. Experiments with $L_h = 500, 600$, and 700 km showed stable performance over a period of 1 year, until the end of 2017. Notably, data assimilation improved not only the observed variable, temperature, but also the unobserved variables, such as geopotential height. This indicates that observation information was propagated to unobserved variables through the data assimilation cycle. In contrast, the experiment with $L_h = 800$ km exhibited filter divergence after September 2017 due to erroneous error covariance associated with the larger localization scale. In addition, the experiment with $L_h = 400$ km kept reducing the RMSEs over a year, but the RMSEs were still higher than those of the other exper-

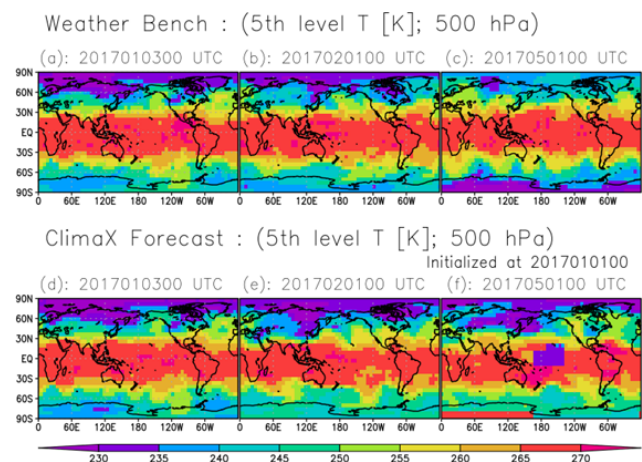


Figure 3. Spatial patterns of temperature (K) at the fifth model level (500 hPa). Panels (a)–(c) are WeatherBench data. Panels (d)–(f) are forecasts by ClimaX initialized at 00:00 UTC on 1 January 2017. Panels (a) and (d) show 00:00 UTC on 3 January 2017, (b) and (e) show 00:00 UTC on 1 February 2017, and (c) and (f) show 00:00 UTC on 1 May 2017.

iments, with the exception of $L_h = 800$ km. This indicates that a localization scale that is too small is suboptimal. This implies that ensemble-based error covariance is beneficial to some extent for propagating the impacts of assimilated observations for distant grid points.

Figure 5 shows the global mean RMSEs for zonal wind, meridional wind, temperature, specific humidity, geopotential height, and surface pressure as a function of the horizontal localization scales averaged over July–December 2017. At smaller localization scales ($L_h = 400$ and 500 km), the analysis RMSEs tended to be lower than the first-guess RMSEs, which suggests that data assimilation was beneficial

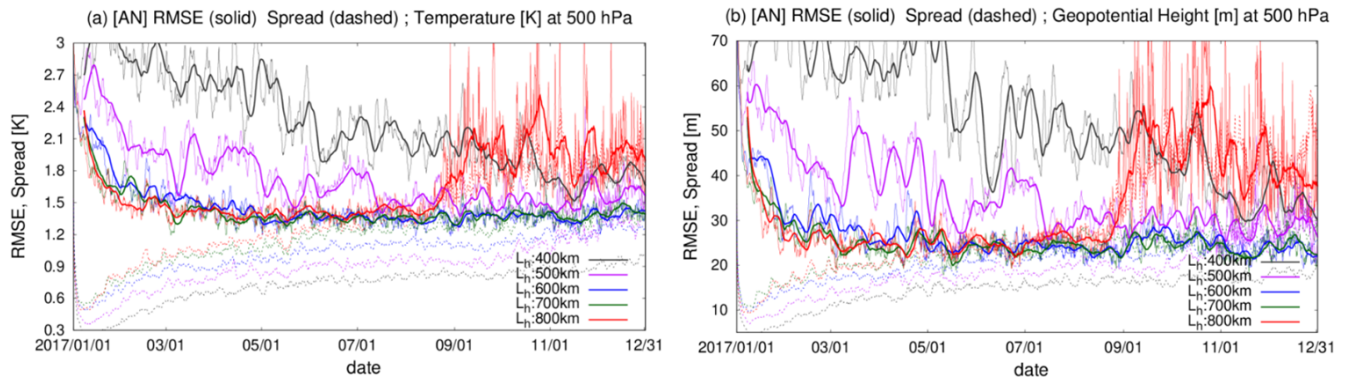


Figure 4. Time series of global mean root mean square errors (RMSEs) verified against WeatherBench data and ensemble spreads for (a) temperature (K) and geopotential height (m) at the fifth model level (500 hPa). The thin and bold solid lines indicate 6-hourly RMSEs and their 7 d running means, respectively. The dashed lines indicate ensemble spreads. The black, purple, blue, green, and red lines indicate the ClimaX–LETKF experiments at localization scales of $L_h = 400, 500, 600, 700,$ and 800 km. The x axis indicates the date (month/day) in 2017.

in reducing errors. Conversely, at larger localization scales ($L_h = 800$ km), analysis RMSEs tended to be higher than the first-guess RMSEs, indicating that data assimilation degraded the analysis, presumably also due to excessive error covariance at larger localization scales. In addition, the analysis RMSEs were slightly higher than the first-guess RMSEs for some variables at $L_h = 700$ km, although the data assimilation cycled stably (Fig. 4). In general, stable filters are expected to yield overall RMSE reduction unless the system is non-chaotic. Therefore, these results for $L_h = 700$ km imply that the present ClimaX exhibits weaker chaotic behavior compared to the real atmosphere.

Among the five experiments, a localization scale of $L_h = 600$ km yielded the lowest analysis RMSEs for most variables. Significant analysis error reductions were observed for temperature and surface pressure. However, no clear impacts were observed for zonal and meridional winds. Even slight degradations were detected, implying that spatial and inter-variable error covariance may not be well represented in our ClimaX–LETKF.

Here, we investigate the spatial patterns of the difference between the analysis and first-guess mean absolute errors, which is given by

$$\text{MAE}_{\text{diff}} = \frac{1}{N_t} \sum_t |\bar{x}_t^a - x_t^{\text{WB}}| - |\bar{x}_t^b - x_t^{\text{WB}}|, \quad (5)$$

where N_t is the sample size, and superscript WB represents WeatherBench data. Negative and positive values indicate improvements and degradations due to data assimilation. Figure 6 shows the MAE_{diff} for four variables (zonal wind at 850 hPa, temperature at 700 hPa, geopotential height at 500 hPa, and surface pressure) based on the experiments with the localization scale $L_h = 500$ km, which resulted in RMSE reductions by data assimilation for most of the variables in Fig. 5. General improvements are seen at grid points with observations for zonal wind and temperature (Fig. 6a

and b). However, there were also slight degradations at grid points surrounding observing stations, such as those in the Arctic Ocean and along the US and Japanese coasts. We also see degradations for geopotential height, where temperature and zonal wind degradations are presented (Fig. 6c). These degradations suggest that ensemble-based spatial error covariance was suboptimal in these regions. In contrast, geopotential height and surface pressure generally improved in the Southern Hemisphere (Fig. 6c and d). In particular, improvements are seen even at grid points surrounding observing stations in the Southern Hemisphere. Specifically, using the spatial and inter-variable error covariance based on AI ensemble forecasts was advantageous for geopotential heights and surface pressure in sparsely observed regions.

Another important property is that ClimaX would be less chaotic than dynamical NWP models, as indicated by the estimated inflation factor β diagnosed by observation–space statistics (Fig. 7). In addition, the larger inflation would also indicate greater model error in ClimaX, which requires stronger inflation to account for the model’s imperfection. Compared to our study, Kotsuki et al. (2017) estimated a much smaller inflation factor for a global ensemble data assimilation system using a dynamical model (see Fig. 10a in Kotsuki et al., 2017). For example, the inflation factors in Kotsuki et al. (2017) were at most around 2.0, whereas ClimaX–LETKF required inflation factors exceeding 5.0. Selz and Craig (2023) noted that an AI-based weather prediction model failed to reproduce rapid initial error growth rates, which would prevent it from replicating the butterfly effect as accurately as dynamical NWP models.

4 Discussion and summary

The optimal localization scale was unexpectedly very small in Fig. 5. Kondo and Miyoshi (2016) pointed out that a larger

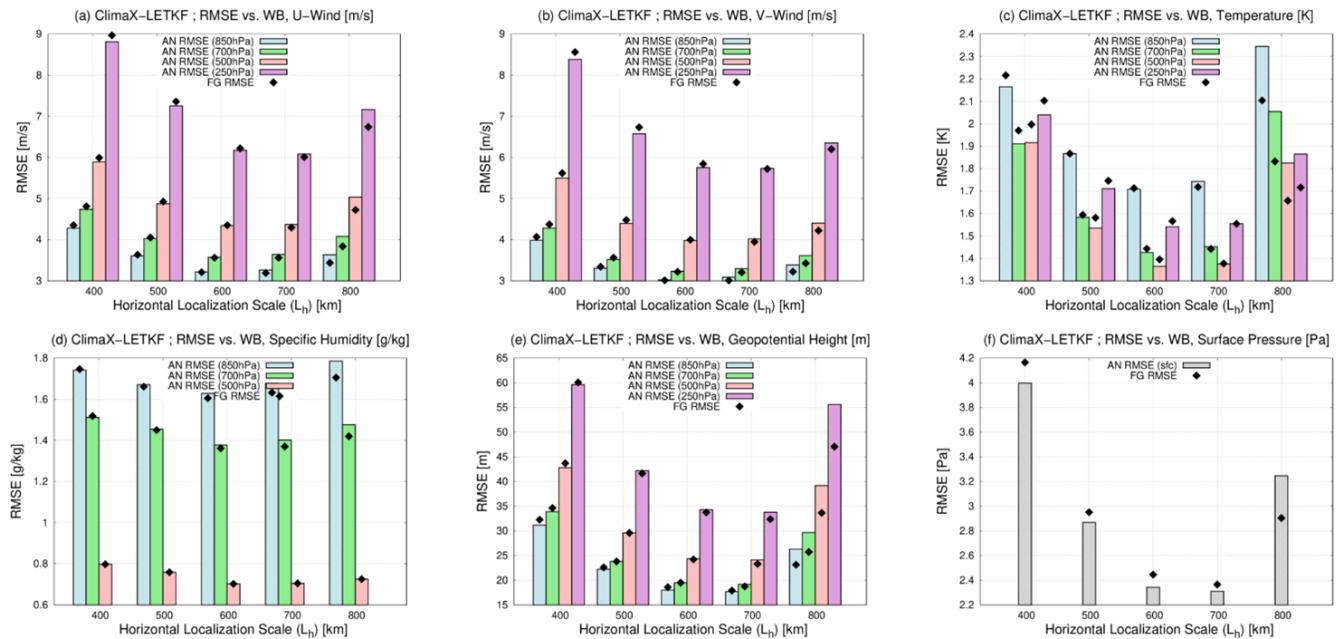


Figure 5. Global mean root mean square errors (RMSEs) verified against WeatherBench (WB) for (a) zonal wind (m s^{-1}), (b) meridional wind (m s^{-1}), (c) temperature (K), (d) specific humidity (g kg^{-1}), (e) geopotential height (m), and (f) surface pressure (hPa) as a function of the horizontal localization scales (km) averaged over July–December 2017. The colored bars and black diamonds indicate analysis (AN) and first-guess (FG) RMSEs, respectively. The blue, green, red, and purple bars in panels (a)–(e) represent the second, third, fifth, and sixth model levels (850, 700, 500, and 250 hPa, respectively). The RMSEs of specific humidity at the sixth model level in panel (d) were too low to be shown.

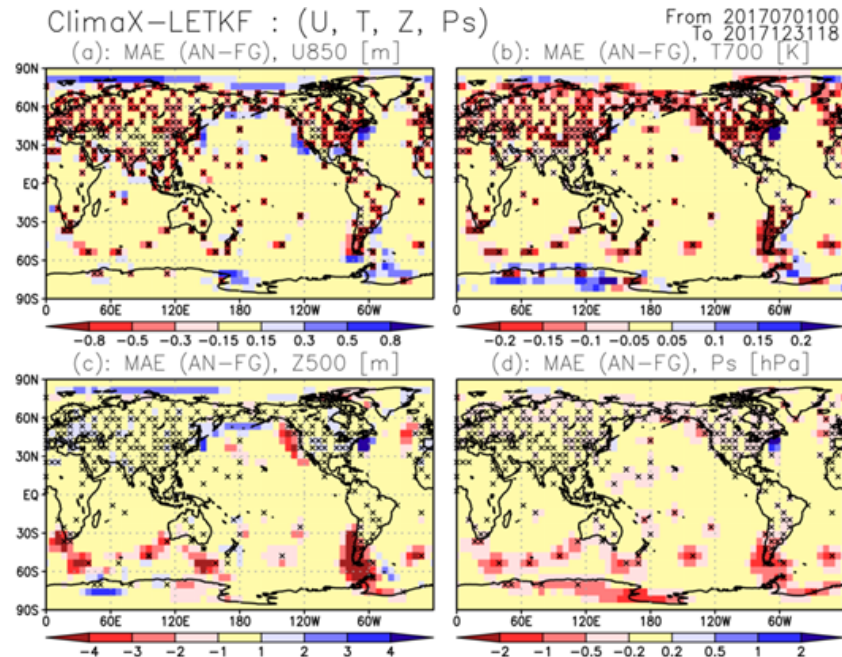


Figure 6. Spatial patterns of difference between analysis (AN) and first-guess (FG) mean absolute errors (MAEs) for (a) zonal wind (m s^{-1}) at 850 hPa, (b) temperature (K) at 700 hPa, (c) geopotential height (m) at 500 hPa, and surface pressure (hPa), averaged over July–December 2017. The warm and cold colors represent improvements and degradations due to data assimilation. Results are for a localization scale of $L_h = 500$ km. The black crosses indicate observing stations.

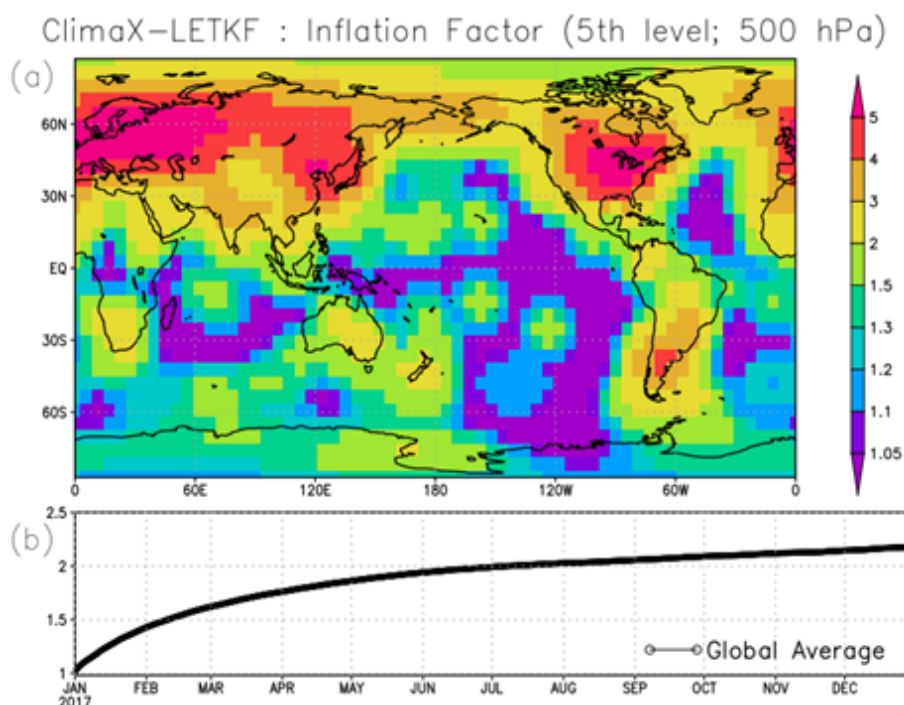


Figure 7. (a) Spatial pattern of the multiplicative inflation factor at the end of the experiment on 18:00 UTC of 31 December 2017. (b) Time series of globally averaged inflation factors. Results are for a localization scale of $L_h = 600$ km.

localization scale is beneficial for low-resolution models and larger ensemble sizes (see Table 1 in Kondo and Miyoshi, 2016). Our optimal localization scale for the 20-member Climax-LETKF was 600 km, which is shorter than the 700–900 km scale of the 20-member LETKF experiments coupled with a dynamical NWP model (also known as SPEEDY; Molteni, 2003) (e.g., Miyoshi and Kondo, 2013, Fig. 2b in Kotsuki and Bishop, 2022). Nevertheless, considering that the SPEEDY model has a finer horizontal resolution (96×48 horizontal grids) than Climax used in this study (64×32 horizontal grids), it remains plausible that Climax captures flow-dependent error covariance less effectively than dynamical NWP models. Bonavita (2024) investigated physical realism of the present AI models (FourCastNet, Pangu-Weather, and GraphCast) and concluded that AI models are not able to properly reproduce sub-synoptic and mesoscale weather phenomena. The suboptimal flow-dependent error covariance in this study can be attributed to physically inconsistent atmospheric fields of the Climax predictions.

Two major advancements are required for AI-based weather prediction models to improve ensemble data assimilation. First, it is imperative that AI models generate physically consistent forecast variables. The accuracy of spatial and inter-variable error covariance would be improved by this enhancement, which would require AI model training procedures to include physical constraints such as hydrostatic and geostrophic balances, in addition to decreasing the mean square errors of the target variables. Second, it is cru-

cial to accurately capture the error growth rate. Our findings demonstrated that error growth based on ensemble Climax predictions was weaker than that of dynamical NWP models, leading to higher inflation factors (Fig. 7). Thus, ensemble forecasts produced by AI weather prediction models likely exhibit insufficient spread. In weather forecasting, capturing forecast uncertainty is as important as providing accurate forecasts. Recent studies have begun to develop models for generating statistically accurate ensembles by using generative models (Price et al., 2025) or by training on probabilistic cost functions (Kochkov et al., 2024). Other possible solutions for improving the error growth are to develop a set of slightly different AI models by randomizing the seed in the AI training process as an analogy to stochastic parameterization (Weyn et al., 2021) or to incorporate the Lyapunov exponent within the cost function of model training (Platt et al., 2023). Note that the present experiments were conducted at a coarse resolution of 5.625° , which may limit the ability of the Climax-LETKF system to accurately diagnose localized weather phenomena. At higher spatial resolutions, AI models may capture mesoscale and sub-synoptic features, potentially leading to more realistic ensemble-based error covariances. Future work will explore the data assimilation system's behaviors at higher resolutions using more advanced versions of AI models with denser observation datasets.

Despite the need for further improvements, this study represents a significant step toward ensemble data assimilation for AI-based weather prediction models. Notably, we demon-

strated that the data assimilation cycled stably for the AI-based weather prediction model ClimaX with the LETKF using covariance inflation and localization techniques. In addition, the ensemble-based error covariance was estimated reasonably well by the AI-based weather prediction model in sparsely observed regions.

Additional research is anticipated for areas identified as requiring further improvements. For that purpose, ensemble data assimilation is a useful tool for diagnosing AI-based weather forecasting models. That is, investigating optimal localization scales, ensemble-based error covariance, and necessary inflation factors gives beneficial insights into understanding the properties of AI models. After achieving these two major advancements, it is important to employ a systematic sensitivity analysis for the localization radius and ensemble size. A suitable inflation method for AI-based weather prediction models also remains to be explored. Comparing the EnKF with variational or ensemble–variational approaches would be an important topic for future investigation. Since AI models require much lower computational costs compared to dynamical NWP models, extending the present study to large-ensemble EnKFs or a local particle filter (LPF) is also an important subject of future studies. Our future work will investigate the applicability of the proposed system to real-time forecasting with higher-resolution AI models with real weather observations, such as PREP-BUFR and satellite radiances. The analysis fields and ensemble spreads generated by the ensemble data assimilation with assimilation of real observations may be applicable to subsequent training of AI models. Most current AI weather models are trained on reanalysis data such as WeatherBench, without explicitly accounting for the uncertainty in analysis (i.e., analysis ensemble spread). By using ensemble spreads or individual ensemble members, the training process of AI models could be improved, such as by relaxing penalties in regions with large ensemble spread.

Beyond weather prediction, data assimilation has been successfully combined with machine-learning-based surrogate models in various fields, including oceanography, hydrology, and wildfire (e.g., Brajard et al., 2021; Cheng et al., 2022; Jeong et al., 2024). It would be beneficial to explore how the EnKF could be applied to diagnose AI-based models in other fields.

Code and data availability. The data assimilation system, experimental data, and visualization scripts used in this paper are archived on Zenodo (<https://doi.org/10.5281/zenodo.13884167>, Kotsuki, 2024a). The original ClimaX version 0.3.1 and LETKF codes are also archived on Zenodo (ClimaX version 0.3.1, <https://doi.org/10.5281/zenodo.14258099>, Kotsuki, 2024b; LETKF, <https://doi.org/10.5281/zenodo.14258014>, Kotsuki, 2024c).

Author contributions. SK developed the ClimaX–LETKF system and employed data assimilation experiments, KS updated and trained ClimaX, and AO contributed to the discussion about the analyses of the experiments.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

Acknowledgements. The authors thank the two reviewers for their constructive comments.

Financial support. This research has been supported by JST Moonshot R&D (JPMJMS2389), the Japan Aerospace Exploration Agency (JAXA) Precipitation Measuring Mission (PMM) 70 (ER4GPF019), the Japan Society for the Promotion of Science (JSPS) (KAKENHI grant nos. JP21H04571, JP21H05002, JP22K18821, and JP25H00752), and the IAAR Research Support Program and VL Program of Chiba University.

Review statement. This paper was edited by Yuefei Zeng and reviewed by two anonymous referees.

References

- Adrian, M., Sanz-Alonso, D., and Willett, R.: Data Assimilation with Machine Learning Surrogate Models: A Case Study with FourCastNet, *Artif. Intell. Earth Syst.*, 4, e240050, <https://doi.org/10.1175/AIES-D-24-0050.1>, 2025.
- Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q.: Accurate medium-range global weather forecasting with 3D neural networks, *Nature*, 619, 533–538, <https://doi.org/10.1038/s41586-023-06545-z>, 2023.
- Bocquet, M., Farchi, A., Finn, T. S., Durand, C., Cheng, S., Chen, Y., Pasmans, I., and Carrassi, A.: Accurate deep learning-based filtering for chaotic dynamics by identifying instabilities without an ensemble, *Chaos*, 34, 091104, <https://doi.org/10.1063/5.0230837>, 2024.
- Bodnar, C., Bruinsma, W. P., Lucic, A., Stanley, M., Allen, A., Brandstetter, J., Garvan, P., Riechert, M., Weyn, J. A., Dong, H., Gupta, J. K., Thambiratnam, K., Archibald, A. T., Wu, C.-C., Heider, E., Welling, M., Turner, R. E., and Perdikaris, P.: A foundation model for the Earth System, *Nature*, 641, 1180–1187, <https://doi.org/10.1038/s41586-025-09005-y>, 2025.

- Bonavita, M.: On some limitations of current machine learning weather prediction models, *Geophys. Res. Lett.*, 51, e2023GL107377, <https://doi.org/10.1029/2023GL107377>, 2024.
- Bonev, B., Kurth, T., Hundt, C., Pathak, J., Baust, M., Kashinath, K., and Anandkumar, A.: Spherical fourier neural operators: Learning stable dynamics on the sphere, in: *Proceedings of the 40th International Conference on Machine Learning*, the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA, 23–29 July 2023, 2806–2823, <https://proceedings.mlr.press/v202/bonev23a.html> (last access: 5 September 2025), 2023.
- Brajard, J., Carrassi, A., Bocquet, M., and Bertino, L.: Combining data assimilation and machine learning to infer unresolved scale parametrization, *Philos. T. R. Soc. A*, 379, 20200086, <https://doi.org/10.1098/rsta.2020.0086>, 2021.
- Chattopadhyay, A., Mustafa, M., Hassanzadeh, P., Bach, E., and Kashinath, K.: Towards physics-inspired data-driven weather forecasting: integrating data assimilation with a deep spatial-transformer-based U-NET in a case study with ERA5, *Geosci. Model Dev.*, 15, 2221–2237, <https://doi.org/10.5194/gmd-15-2221-2022>, 2022.
- Chattopadhyay, A., Nabizadeh, E., Bach, E., and Hassanzadeh, P.: Deep learning-enhanced ensemble-based data assimilation for high-dimensional nonlinear dynamical systems, *J. Comput. Phys.*, 477, 111918, <https://doi.org/10.1016/j.jcp.2023.111918>, 2023.
- Chen, K., Bai, L., Ling, F., Ye, P., Chen, T., Luo, J.-J., Chen, H., Xiao, Y., Chen, K., Han, T., and Ouyang, W.: Towards an end-to-end artificial intelligence driven global weather forecasting system, *arXiv [preprint]*, <https://doi.org/10.48550/arXiv.2312.12462>, 8 April 2024.
- Cheng, S., Prentice, I. C., Huang, Y., Jin, Y., Guo, Y. K., and Arcucci, R.: Data-driven surrogate model with latent data assimilation: Application to wildfire forecasting, *J. Comput. Phys.*, 464, 111302, <https://doi.org/10.1016/j.jcp.2022.111302>, 2022.
- Desroziers, G., Berre, L., Chapnik, B., and Poli, P.: Diagnosis of observation, background and analysis-error statistics in observation space, *Q. J. Roy. Meteor. Soc.*, 131, 3385–3396, <https://doi.org/10.1256/qj.05.108>, 2005.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houshy, N.: An image is worth 16x16 words: Transformers for image recognition at scale, *The Ninth International Conference on Learning Representations (ICLR2021)*, online, 3–7 May 2021, <https://openreview.net/forum?id=YicbFdNTTy> (last access: 5 September 2025), 2021.
- Jeong, M., Kwon, M., Cha, J. H., and Kim, D. H.: High flow prediction model integrating physically and deep learning based approaches with quasi real-time watershed data assimilation, *J. Hydrol.*, 636, 131304, <https://doi.org/10.1016/j.jhydrol.2024.131304>, 2024.
- Jonkman, S. N., Curran, A., and Bouwer, L. M.: Floods have become less deadly: an analysis of global flood fatalities 1975–2022, *Nat. Hazards*, 120, 6327–6342, <https://doi.org/10.1007/s11069-024-06444-0>, 2024.
- Hamilton, F., Berry, T., and Sauer, T.: Ensemble Kalman filtering without a model, *Phys. Rev. X*, 6, 011021, <https://doi.org/10.1103/PhysRevX.6.011021>, 2016.
- Houtekamer, P. L., and Zhang, F.: Review of the ensemble Kalman filter for atmospheric data assimilation, *Mon. Wea. Rev.*, 144, 4489–4532, <https://doi.org/10.1175/MWR-D-15-0440.1>, 2016.
- Hunt, B. R., Kostelich, E. J., and Szunyogh, I.: Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter, *Physica D*, 230, 112–126, <https://doi.org/10.1016/j.physd.2006.11.008>, 2007.
- Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., Klöwer, M., Lottes, J., Rasp, S., Düben, P., Hatfield, S., Battaglia, P., Sanchez-Gonzalez, A., Willson, M., Brenner, M. P., and Hoyer, S.: Neural general circulation models for weather and climate, *Nature*, 632, 1060–1066, <https://doi.org/10.1038/s41586-024-07744-y>, 2024.
- Kondo, K. and Miyoshi, T.: Impact of removing covariance localization in an ensemble Kalman filter: Experiments with 10 240 members using an intermediate AGCM, *Mon. Weather Rev.*, 144, 4849–4865, <https://doi.org/10.1175/MWR-D-15-0388.1>, 2016.
- Kotsuki, S.: Experimental data, source codes and scripts used in Kotsuki et al. (2024) submitted to GMD, Zenodo [data set], <https://doi.org/10.5281/zenodo.13884167>, 2024a.
- Kotsuki, S.: Original source code of the ClimaX version 0.3.1 used in Kotsuki et al. (2024) submitted to GMD, Zenodo [code], <https://doi.org/10.5281/zenodo.14258099>, 2024b.
- Kotsuki, S.: Original source code of the LETKF used in Kotsuki et al. (2024) submitted to GMD, Zenodo [code], <https://doi.org/10.5281/zenodo.14258014>, 2024c.
- Kotsuki, S. and Bishop, H. C.: Implementing Hybrid Background Error Covariance into the LETKF with Attenuation-based Localization: Experiments with a Simplified AGCM, *Mon. Weather Rev.*, 150, 283–302, <https://doi.org/10.1175/MWR-D-21-0174.1>, 2022.
- Kotsuki, S., Ota, Y., and Miyoshi, T.: Adaptive covariance relaxation methods for ensemble data assimilation: experiments in the real atmosphere, *Q. J. Roy. Meteor. Soc.*, 143, 2001–2015, <https://doi.org/10.1002/qj.3060>, 2017.
- Kotsuki, S., Miyoshi, T., Kondo, K., and Potthast, R.: A local particle filter and its Gaussian mixture extension implemented with minor modifications to the LETKF, *Geosci. Model Dev.*, 15, 8325–8348, <https://doi.org/10.5194/gmd-15-8325-2022>, 2022.
- Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirmsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Meroze, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., and Battaglia, P.: Learning skillful medium-range global weather forecasting, *Science*, 382, 1416–1421, <https://doi.org/10.1126/science.adi2336>, 2023.
- Luk, E., Bach, E., Baptista, R., and Stuart, A.: Learning optimal filters using variational inference, *arXiv [preprint]*, <https://doi.org/10.48550/arXiv.2406.18066>, 26 June 2024.
- McCabe, M. and Brown, J.: Learning to assimilate in chaotic dynamical systems, in: *Advances in Neural Information Processing Systems 34*, 35th Conference on Neural Information Processing Systems (NeurIPS 2021), online, 6–14 December 2021, 12237–12250, <https://openreview.net/forum?id=ctusEbqyLwO> (last access: 5 September 2025), 2021.
- Miyoshi, T.: The Gaussian Approach to Adaptive Covariance Inflation and Its Implementation with the Local Ensemble Transform Kalman Filter, *Mon. Weather Rev.*, 139, 1519–1535, <https://doi.org/10.1175/2010MWR3570.1>, 2011.

- Miyoshi, T. and Kondo, K.: A multi-scale localization approach to an ensemble Kalman filter, *SOLA*, 9, 170–173, <https://doi.org/10.2151/sola.2013-038>, 2013.
- Molteni, F.: Atmospheric simulations using a GCM with simplified physical parametrizations. I: Model climatology and variability in multi-decadal experiments, *Clim. Dynam.*, 20, 175–191, <https://doi.org/10.1007/s00382-002-0268-2>, 2003.
- Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K., and Grover, A.: ClimaX: A foundation model for weather and climate, *arXiv [preprint]*, <https://doi.org/10.48550/arXiv.2301.10343>, 18 December 2023.
- Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z., Azizzadehsheli, K., Hassanzadeh, P., Kashinath, K., and Anandkumar, A.: Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators, *arXiv [preprint]*, <https://doi.org/10.48550/arXiv.2202.11214>, 22 February 2022.
- Penny, S. G., Smith, T. A., Chen, T. C., Platt, J. A., Lin, H. Y., Goodliff, M., and Abarbanel, H. D.: Integrating recurrent neural networks with data assimilation for scalable data-driven state estimation, *J. Adv. Model. Earth Sy.*, 14, e2021MS002843, <https://doi.org/10.1029/2021MS002843>, 2022.
- Platt, J. A., Penny, S. G., Smith, T. A., Chen, T. C., and Abarbanel, H. D.: Constraining chaos: Enforcing dynamical invariants in the training of reservoir computers, *Chaos*, 33, 103107, <https://doi.org/10.1063/5.0156999>, 2023.
- Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., Ewalds, T., Stott, J., Mohamed, S., Battaglia, P., Lam, R., and Willson, M.: Probabilistic weather forecasting with machine learning, *Nature*, 637, 84–90, <https://doi.org/10.1038/s41586-024-08252-9>, 2025.
- Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., and Thuerey, N.: WeatherBench: a benchmark data set for data-driven weather forecasting, *J. Adv. Model. Earth Sy.*, 12, e2020MS002203, <https://doi.org/10.1029/2020MS002203>, 2020.
- Rasp, S., Hoyer, S., Meroze, A., Langmore, I., Battaglia, P., Russell, T., Sanchez-Gonzalez, A., Yang, V., Carver, R., Agrawal, S., Chantry, M., Ben Bouallegue, Z., Dueben, P., Bromberg, C., Sisk, J., Barrington, L., Bell, A., and Sha, F.: WeatherBench 2: A benchmark for the next generation of data-driven global weather models, *J. Adv. Model. Earth Sy.*, 16, e2023MS004019, <https://doi.org/10.1029/2023MS004019>, 2024.
- Selz, T. and Craig, G. C.: Can artificial intelligence-based weather prediction models simulate the butterfly effect?, *Geophys. Res. Lett.*, 50, e2023GL105747, <https://doi.org/10.1029/2023GL105747>, 2023.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I.: Attention Is All You Need, *arXiv [preprint]*, <https://doi.org/10.48550/arXiv.1706.03762>, 2 June 2017.
- Vaughan, A., Markou, S., Tebbutt, W., Requeima, J., Bruinsma, W. P., Andersson, T. R., Herzog, M., Lane, N. D., Chantry, M., Hosking, J. S., and Turner, R. E.: Aardvark Weather: end-to-end data-driven weather forecasting, *arXiv [preprint]*, <https://doi.org/10.48550/arXiv.2404.00411>, 13 July 2024.
- Weyn, J. A., Durran, D. R., Caruana, R., and Cresswell-Clay, N.: Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models, *J. Adv. Model. Earth Sy.*, 13, e2021MS002502, <https://doi.org/10.1029/2021MS002502>, 2021.
- World Economic Forum: The global risks report 2023 18th Edition, World Economic Forum, https://www3.weforum.org/docs/WEF_Global_Risks_Report_2023.pdf (last access: 05 September 2025), 2023.
- Xiao, Y., Bai, L., Xue, W., Chen, K., Han, T., and Ouyang, W.: Fengwu-4dvar: Coupling the data-driven weather forecasting model with 4d variational assimilation, *arXiv [preprint]*, <https://doi.org/10.48550/arXiv.2312.12455>, 19 May 2024.