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

Shunji Kotsuki[1,2,3], Kenta Shiraishi[4], and Atsushi Okazaki[1,2]

[1]Institute for Advanced Academic Research, Chiba University, Chiba, Japan
[2]Center for Environmental Remote Sensing, Chiba University, Chiba, Japan
[3]Research Institute of Disaster Medicine, Chiba University, Chiba, Japan
[4]Graduate School of Science and Engineering, Chiba University, Chiba, Japan

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

Abstract.

Artificial intelligence (AI)-based weather prediction research is growing rapidly and has shown to be competitive with the advanced dynamic numerical weather prediction 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 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 competitive to dynamical NWP models, FourCastNet, 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. 2024). These machine learning approaches have been shown to be competitive with state-of-the-art NWP models (e.g., Kochkov et al. 2024). Progresses 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, Dosovitski et al. 2020). The ViT, which has been explored in language models and image classifications, was 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. (2023) and Adrian et al. (2024). On the other hand, recent studies have started to solve the inverse problem inherent in data assimilation by deep neural networks (McCabe and Brown 2021, Chen et al. 2023, Boucquet et al. 2024, Luk et al. 2024, Vaughan et al. 2024). There have been some studies employing 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 for assessing 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), Deutscher Wetterdienst (DWD) and Japan Meteorological Agency (JMA). Using the coupled ClimaX–LETKF data assimilation system, we investigated several key aspects of AI-based weather prediction model, including whether the data assimilation cycles stably for the ClimaX AI weather prediction model using ensemble Kalman filters; 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 update of the AI weather models with meteorological observations.

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

2 Methods and experiments

2.1 ClimaX Model

The ClimaX (Nguyen et al. 2023) is a ViT-based AI weather prediction model for the global atmosphere. Variable tokenization and variable aggregation are the key components of the ClimaX architecture upon 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° × 5.625°. 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 the ClimaX entirely with the additional outputs (i.e., no transfer learning). It took approximately 4 hours with four GPU of NVIDIA RTX 6000Ada. Anomaly correlation coefficients increased and root mean square errors (RMSEs) decreased in Figure 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 $\mathbf{X}_t \equiv \left\{\mathbf{x}_t^{(1)}, \ldots, \mathbf{x}_t^{(m)}\right\}$ be an ensemble state matrix, whose ensemble mean and perturbation is given by $\bar{\mathbf{x}}_t$ $(\in \mathbb{R}^n)$ and $\delta\mathbf{X}_t \equiv \left\{\mathbf{x}_t^{(1)} - \bar{\mathbf{x}}_t, \ldots, \mathbf{x}_t^{(m)} - \bar{\mathbf{x}}_t\right\}$ $(\in \mathbb{R}^{n\times m})$, respectively. Here, n and m are the system and ensemble sizes. The superscript (i) and subscript t denote the ith ensemble member and indicates the time, respectively. The EnKFs, including LETKF, estimate error covariance $\mathbf{P}$ $(\in \mathbb{R}^{n\times n})$ according to sample estimates based on ensemble perturbation:

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

The analysis update equation of the LETKF is given by:

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

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

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 linear observation operator matrix, and $\mathbf{1}$ is a row vector whose all elements are 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)'s 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 **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 realizes the localization (Hunt et al. 2007) whose function is given by:

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

where l is the localization function, and its inverse l^{-1} is multiplied to inflate **R** for the localization. Horizontal and vertical distances (km and log(Pa)) from 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 was 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 (Figure 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 a year of 2017, which is not used for training and validation of the ClimaX. The ensemble size is 20. Their initial conditions for 00 UTC January 1 in 2017 were taken from WeatherBench data in 2006, which were sampled every 12 hours from 00 UTC January 1 in 2006. Data assimilation experimental results were verified against WeatherBench data.

It should be noted that we were unable to conduct observation system simulation experiments (k.a. OSSEs), which requires a natural run by ClimaX. This is because ClimaX could not produce long-term forecasts within our experimental configurations. A typical example is shown in Figure 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 AIs cannot 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 the ClimaX to synchronize with the real atmosphere as shown in section 3. This property in AI models were theoretically demonstrated by Adrian et al. (2024), which 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 one 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 are still higher than those of the other experiments with the exception of L_h = 800 km. It indicates that a too small localization scale is suboptimal. This implies that ensemble-based error covariance is beneficial to some extent to propagating the impacts of assimilated observation 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 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 (Figure 4). In general, stable filters are expected to yield overall RMSE reduction unless the system is non-chaotic. Therefore, this results for L_h = 700 km imply that the present ClimaX exhibits weaker chaotic behaviour 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:

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

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, respectively. 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 variables in Figure 5. General improvements are seen at grid points with observations for zonal wind and temperature (Figs. 6 a and b). However, there were also slight degradations at grid points surrounding observing stations, such as those in 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. 6 c). These degradations suggest ensemble-based spatial error covariance were suboptimal in these regions. In contrast, geopotential height and surface pressure generally improved in the Southern Hemisphere (Figs. 6 c 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-based ensemble forecasts was advantageous for geopotential heights and surface pressure in sparsely observed regions.

Another important property is that the ClimaX would be less chaotic than dynamical NWP models, as indicated by the estimated inflation factor β diagnosed by observation-space statistics (Figure 7). In addition, the larger inflation would also indicate greater model error in ClimaX, which requires stronger inflation to account for model's imperfection. Compared to our study, Kotsuki et al. (2017) estimated much smaller inflation factor for a global ensemble data assimilation system using a dynamical model (cf. Fig. 10a in Kotsuki et al. 2017). For example, the inflation factors in Kotsuki et al. (2017) were at most around 2.0, whereas the ClimaX-LETKF required inflation factors exceeding 5.0. Selz and Craig (2022) 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 very small unexpectedly in Figure 5. Kondo and Miyoshi (2016) pointed out that a larger localization scale is beneficial for low-resolution models and larger ensemble sizes (cf. 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., Kondo and Miyoshi 2013, Figure 2b in Kotsuki and Bishop 2022). Nevertheless, considering that the SPEEDY model has a finer horizontal resolution (96 × 48 horizontal grids) than the 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 physical 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 the hydrostatic and geostrophic balances, in addition to decreasing the mean square errors of the target variables. Second, it is crucial to accurately capture error growth rate. Our findings demonstrated that error growth based on ensemble ClimaX predictions were weaker than those of dynamical NWP models, leading to higher inflation factors (Figure 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., 2024) or by training on probabilistic cost functions (Kochkov et al., 2024). Other possible solutions for improving the error growth is to develop a set of slightly different AI models by randomizing the seed in the AI training process as an analogy of stochastic parameterization (Weyn et al. 2021) or to incorporate 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 behaviours 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 demonstrated 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 reasonably estimated 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. Namely, investigating optimal localization scales, ensemble-based error covariance and necessary inflation factors give beneficial insights to understand properties of AI models. After achieving the two major advancements, it is important to employ a systematic sensitivity analysis for 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 LPF is also important subjects of future studies. Our future work will

investigate the applicability of the proposed system for real-time forecasting with higher-resolution AI models with real weather observations such as PREPBUFR and satellite radiances. The analysis fields and ensemble spreads generated by 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, 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 manuscript are archived on Zonodo (https://zenodo.org/records/13884167; doi: 10.5281/zenodo.13884167). The original ClimaX version 0.3.1 and LETKF codes are also archived on Zenodo; ClimaX version 0.3.1 (https://zenodo.org/records/14258100, doi: 10.5281/zenodo.14258099) and LETKF (https://zenodo.org/records/14258014, doi: 10.5281/zenodo.14258014).

Author contributions

SK developed the ClimaX-LETKF system and employed data assimilation experiments, KS updated and trained the ClimaX, and AO made large amount of discussion about the analyses of the experiments.

Competing interests

The authors have no competing interests to declare.

Acknowledgements

This study was partly supported by the JST Moonshot R&D (JPMJMS2389), the Japan Aerospace Exploration Agency (JAXA) Precipitation Measuring Mission (PMM) (ER4GPF019), the Japan Society for the Promotion of Science (JSPS) KAKENHI grants JP21H04571, JP21H05002, JP22K18821, JP25H00752, and the IAAR Research Support Program and VL Program of Chiba University.

References

Adrian, M., Sanz-Alonso, D., and Willett, R. (2024): Data Assimilation with Machine Learning Surrogate Models: A Case Study with FourCastNet. arXiv preprint arXiv:2405.13180.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q. (2023): Accurate medium-range global weather forecasting with 3D neural networks. Nature, 619(7970), 533-538. doi: 10.1038/s41586-023-06545-z

Bodnar, C., and Coauthors (2024): Aurora: A foundation model of the atmosphere. arXiv preprint arXiv:2405.13063.

Bonavita, M. (2024): On some limitations of current machine learning weather prediction models. Geophys. Res. Lett., 51(12), e2023GL107377. doi: 10.1029/2023GL107377

Bonev, B., Kurth, T., Hundt, C., Pathak, J., Baust, M., Kashinath, K., and Anandkumar, A. (2023): Spherical fourier neural operators: Learning stable dynamics on the sphere. In International conference on machine learning (pp. 2806-2823). PMLR.

Bocquet, M., Farchi, A., Finn, T. S., Durand, C., Cheng, S., Chen, Y., Pasmans, I. and Carrassi, A. (2024): Accurate deep learning-based filtering for chaotic dynamics by identifying instabilities without an ensemble. Chaos, 34(9). doi: 10.1063/5.0230837

Brajard, J., Carrassi, A., Bocquet, M., and Bertino, L. (2021): Combining data assimilation and machine learning to infer unresolved scale parametrization. Philosophical Transactions of the Royal Society A, 379(2194), 20200086. doi: 10.1098/rsta.2020.0086

Chattopadhyay, A., Mustafa, M., Hassanzadeh, P., Bach, E., and Kashinath, K. (2022): Towards physics-inspired data-driven weather forecasting: Integrating data assimilation with a deep spatial-transformer-based U-NET in a case study with ERA5. Geoscientific Model Development, 15(5), 2221–2237. doi: 10.5194/gmd-15-2221-2022

Chattopadhyay, A., Nabizadeh, E., Bach, E., and Hassanzadeh, P. (2023): Deep learning-enhanced ensemble-based data assimilation for high-dimensional nonlinear dynamical systems. Journal of Computational Physics, 477, 111918. doi:10.1016/j.jcp.2023.111918

Chen, K., and Coauthors (2023): Towards an end-to-end artificial intelligence driven global weather forecasting system. arXiv preprint arXiv:2312.12462.

Cheng, S., Prentice, I. C., Huang, Y., Jin, Y., Guo, Y. K., and Arcucci, R. (2022): Data-driven surrogate model with latent data assimilation: Application to wildfire forecasting. Journal of Computational Physics, 464, 111302. doi: 10.1016/j.jcp.2022.111302

Desroziers G., Berre L., Chapnik B., and Poli P. (2005): Diagnosis of observation, background and analysis-error statistics in observation space. Q. J. R. Meteorol. Soc., 131, 3385–3396. doi:10.1256/qj.05.108

Dosovitskiy, A., and Coauthors (2020): An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Jeong, M., Kwon, M., Cha, J. H., and Kim, D. H. (2024): High flow prediction model integrating physically and deep learning based approaches with quasi real-time watershed data assimilation. Journal of Hydrology, 636, 131304. doi: 10.1016/j.jhydrol.2024.131304

Jonkman, S. N., Curran, A., and Bouwer, L. M. (2024): Floods have become less deadly: an analysis of global flood fatalities 1975–2022. Nat. Hazards, 1-16. doi: 10.1007/s11069-024-06444-0

Hamilton, F., Berry, T., and Sauer, T. (2016): Ensemble Kalman filtering without a model. Physical Review X, 6(1), 011021. doi: 0.1103/PhysRevX.6.011021

Kochkov, D. and Coauthors (2024): Neural general circulation models for weather and climate. Nature, 1-7. doi: 10.1038/s41586-024-07744-y

Kondo, K., and Miyoshi, T. (2016): Impact of removing covariance localization in an ensemble Kalman filter: Experiments with 10 240 members using an intermediate AGCM. Mon. Wea. Rev., 144, 4849–4865. doi: 10.1175/MWR-D-15-0388.1

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, 2024.

Kotsuki, S.: Original source code of the ClimaX version 0.3.1 used in Kotsuki et al. (2024) submitted to GMD, Zenodo [data set], https://doi.org/10.5281/zenodo.14258099, 2024.

Kotsuki, S.: Original source code of the LETKF used in Kotsuki et al. (2024) submitted to GMD, Zenodo [data set], https://doi.org/10.5281/zenodo.14258014, 2024.

Kotsuki, S., Ota. Y., and Miyoshi, T. (2017): Adaptive covariance relaxation methods for ensemble data assimilation : experiments in the real atmosphere, Q. J. R. Meteorol. Soc., 143, 2001–2015. doi: 10.1002/qj.3060

Kotsuki, S., and Bishop, H. C. (2022): Implementing Hybrid Background Error Covariance into the LETKF with Attenuation-based Localization: Experiments with a Simplified AGCM. Mon. Wea. Rev., 150, 283-302. doi: 10.1175/MWR-D-21-0174.1

Kotsuki, S., Miyoshi, T., Kondo K., and Potthast R. (2022): A Local Particle Filter and Its Gaussian Mixture Extension Implemented with Minor Modifications to the LETKF. Geosci. Model Dev., 15, 8325-8348. doi: 10.5194/gmd-15-8325-2022

Lam, R., and Coauthors (2023): Learning skillful medium-range global weather forecasting. Science, 382(6677), 1416-1421. doi: 10.1126/science.adi2336

Luk, E., Bach, E., Baptista, R., and Stuart, A. (2024): Learning optimal filters using variational inference. arXiv preprint arXiv:2406.18066.

McCabe, M., and Brown, J. (2021): Learning to assimilate in chaotic dynamical systems. Advances in neural information processing systems, 34, 12237-12250.

Miyoshi T. (2011): The Gaussian Approach to Adaptive Covariance Inflation and Its Implementation with the Local Ensemble Transform Kalman Filter. Mon. Wea. Rev., 139, 1519–1535. doi:10.1175/2010MWR3570.1

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

Molteni, F. (2003): Atmospheric simulations using a GCM with simplified physical parametrizations. I: Model climatology and variability in multi-decadal experiments. Clim. Dyn., 20, 175–191. doi: 10.1007/s00382-002-0268-2

Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K., and Grover, A. (2023): ClimaX: A foundation model for weather and climate. arXiv preprint arXiv:2301.10343.

Pathak, J., and Coauthors (2022): Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. arXiv preprint arXiv:2202.11214.

Penny, S. G., Smith, T. A., Chen, T. C., Platt, J. A., Lin, H. Y., Goodliff, M., and Abarbanel, H. D. (2022): Integrating recurrent neural networks with data assimilation for scalable data-driven state estimation. J. Adv. Modeling Earth Syst., 14, e2021MS002843. doi: 10.1029/2021MS002843

Platt, J. A., Penny, S. G., Smith, T. A., Chen, T. C., and Abarbanel, H. D. (2023): Constraining chaos: Enforcing dynamical invariants in the training of reservoir computers. Chaos, 33. doi: 10.1063/5.0156999

Price, I., and Coauthors (2025): Probabilistic weather forecasting with machine learning. Nature, 637, 84–90. doi: 10.1038/s41586-024-08252-9

Rasp, S., and Coauthors (2020): WeatherBench: a benchmark data set for data-driven weather forecasting. J. Adv. Modeling Earth Syst., 12(11), e2020MS002203. doi: 10.1029/2020MS002203

Rasp, S., and Coauthors (2024): WeatherBench 2: A benchmark for the next generation of data-driven global weather models. J. Adv. Modeling Earth Syst., 16(6), e2023MS004019. doi: 10.1029/2023MS004019

Selz, T., and Craig, G. C. (2023): Can artificial intelligence-based weather prediction models simulate the butterfly effect?. Geophys. Res. Lett., 50(20), e2023GL105747. doi: 10.1029/2023GL105747

Vaswani, A., and Coauthors (2017): Attention is all you need. Advances in neural information processing systems, 30. doi: 10.1109/ICASSP39728.2021.9413901

Vaughan, A., and Coauthors (2024): Aardvark Weather: end-to-end data-driven weather forecasting. arXiv preprint arXiv:2404.00411.

Weyn, J. A., Durran, D. R., Caruana, R., and Cresswell-Clay, N. (2021): Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models. J. Adv. Modeling Earth Syst., 13(7), e2021MS002502. doi: 10.1029/2021MS002502

World Economic Forum (2023): The global risks report 2023 18th Edition. World Economic Forum. https://www3.weforum.org/docs/WEF_Global_Risks_Report_2023.pdf (last access, July 12, 2024)

Xiao, Y., Bai, L., Xue, W., Chen, K., Han, T., and Ouyang, W. (2023): Fengwu-4dvar: Coupling the data-driven weather forecasting model with 4d variational assimilation. arXiv preprint arXiv:2312.12455.

Tables

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 associated error standard deviation (Error SD).

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

※ Specific humidity is observed up to 4th model level (i.e., 925, 850, 700 and 600 hPa).

Figures

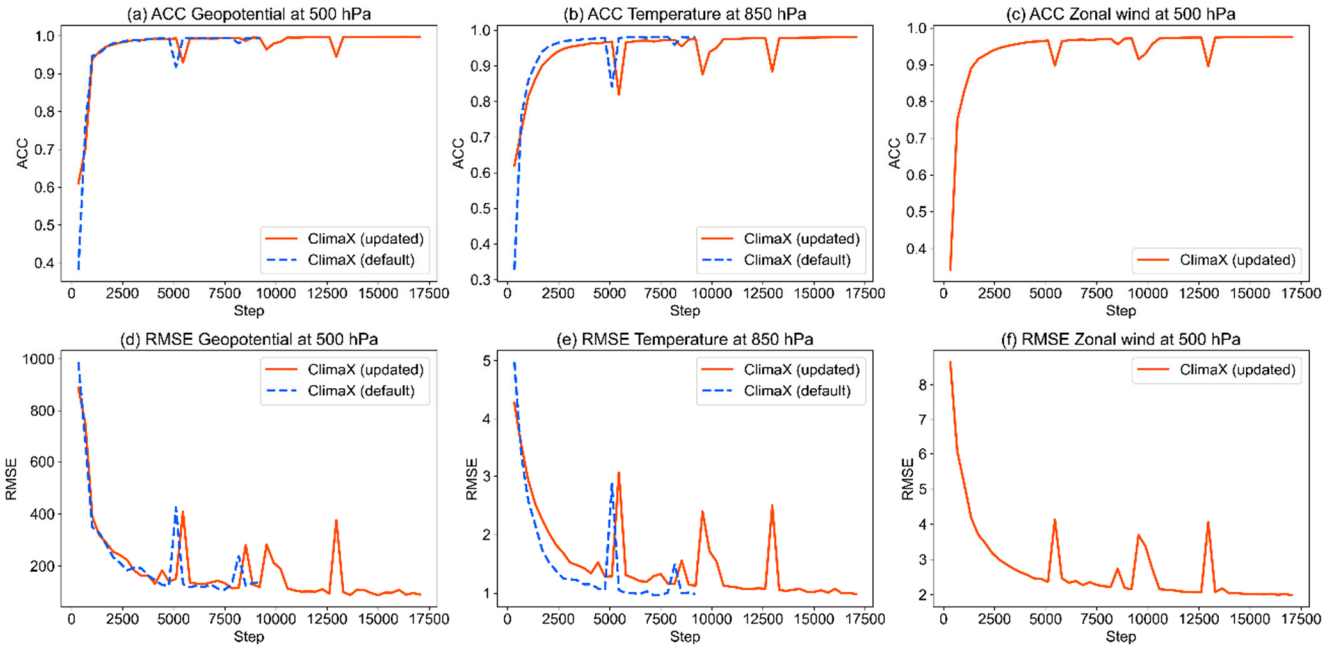


Figure 1: Training curves for the default and updated ClimaX models (dashed blue and solid orange lines) verified against WeatherBench data in 2016, as a function of the number of training steps. Each training step includes 64 training data in a mini batch. Panels (a-c) and (d-f) show anomaly correlation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

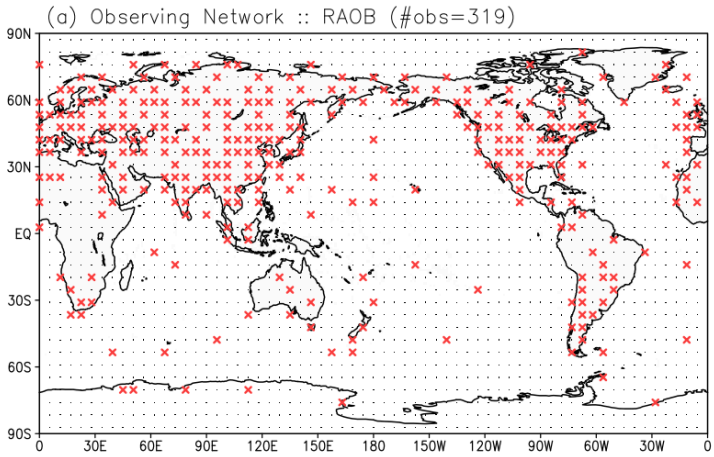


Figure 2. The observing network. Small black dots and red crosses represent model grid points and observing points, respectively.

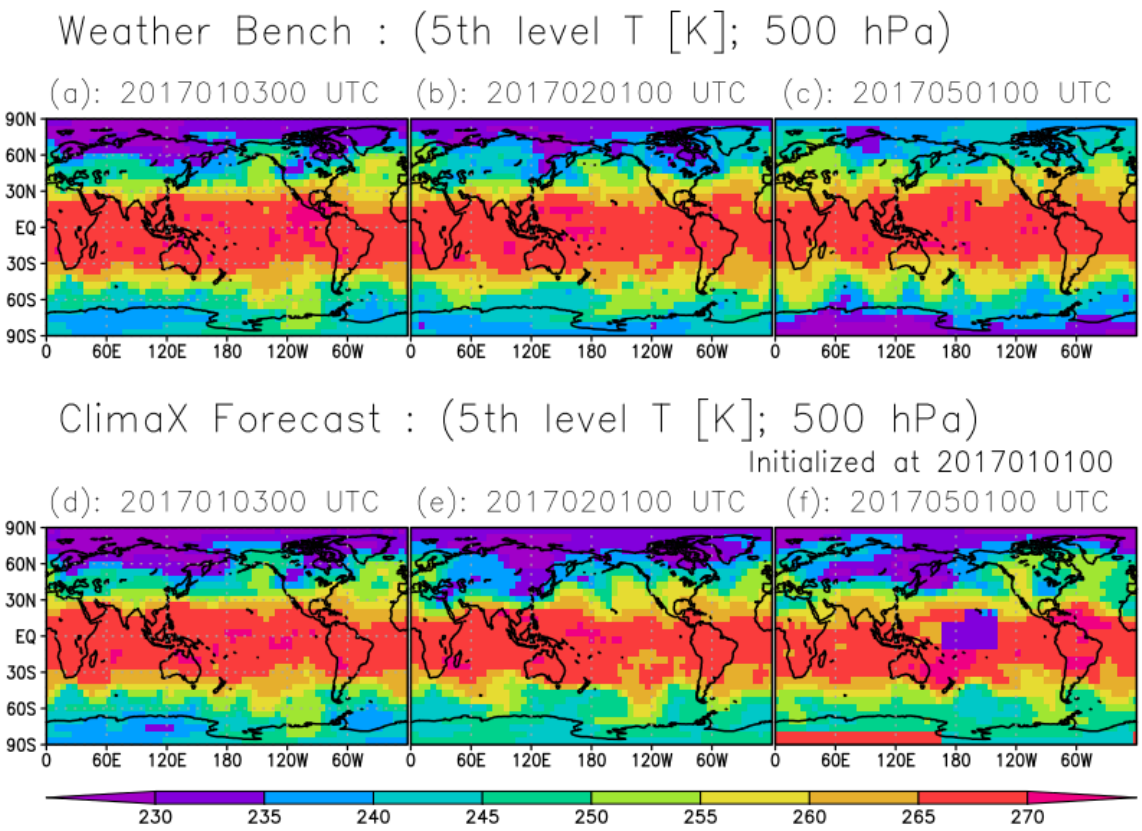


Figure 3: Spatial patterns of temperature (K) at 5th model level (500 hPa). Panels (a-c) are WeatherBench data. Panels (d-f) are forecasts by ClimaX initialized at 0000 UTC of January 1, 2017. Panels (a, d) show 0000 UTC of January 3, 2017, (b, e) show 0000 UTC of February 1, 2017, and (c, f) show 0000 UTC of May 1, 2017, respectively.

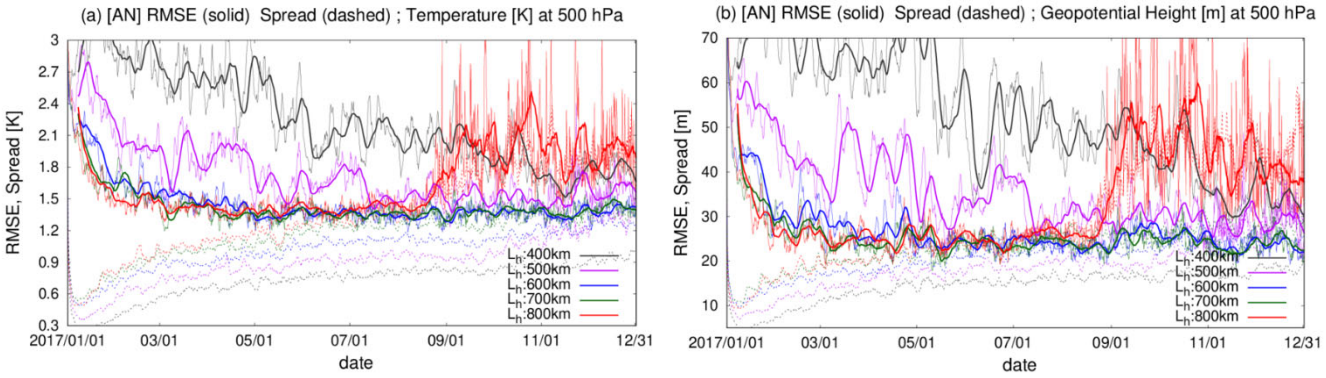


Figure 4: Time series of global-mean root mean square errors (RMSEs) verified against WeatherBench data, and ensemble spreads for (a) temperature (K) and geopotential height (m) at the fifth model level (= 500 hPa). Thin and bold solid lines indicate 6-hourly RMSEs and their 7-day running means, respectively. Dashed lines indicate ensemble spreads. 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 abscissa indicates the date (month/day) in 2017.

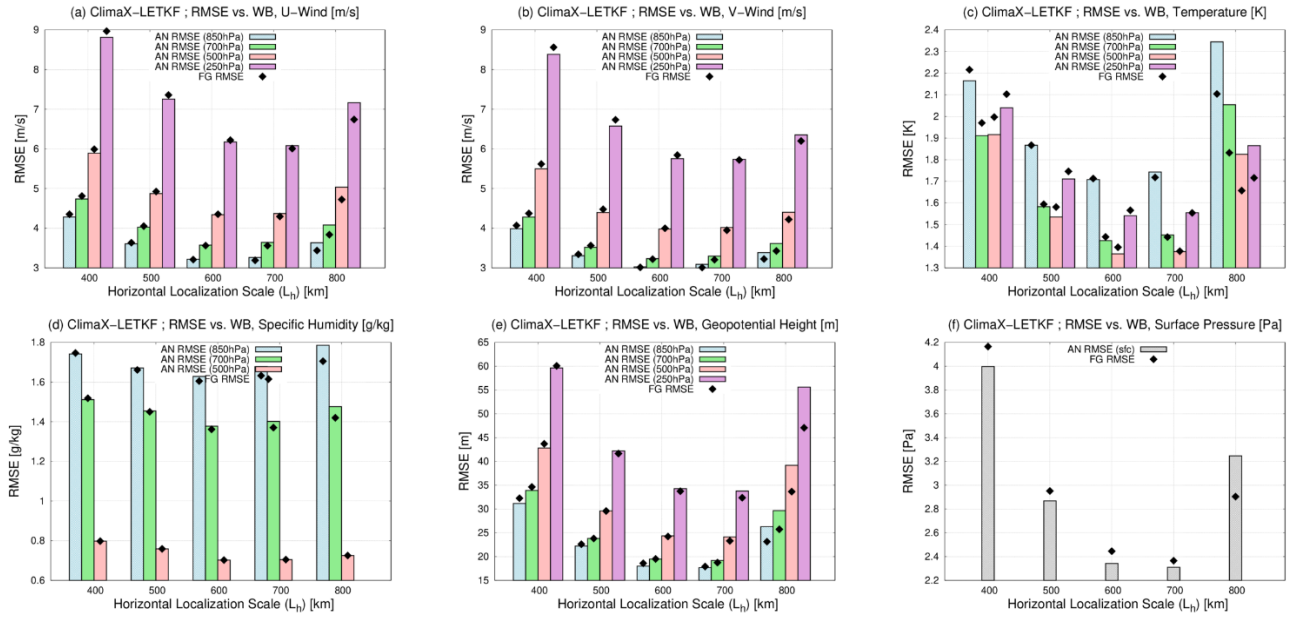
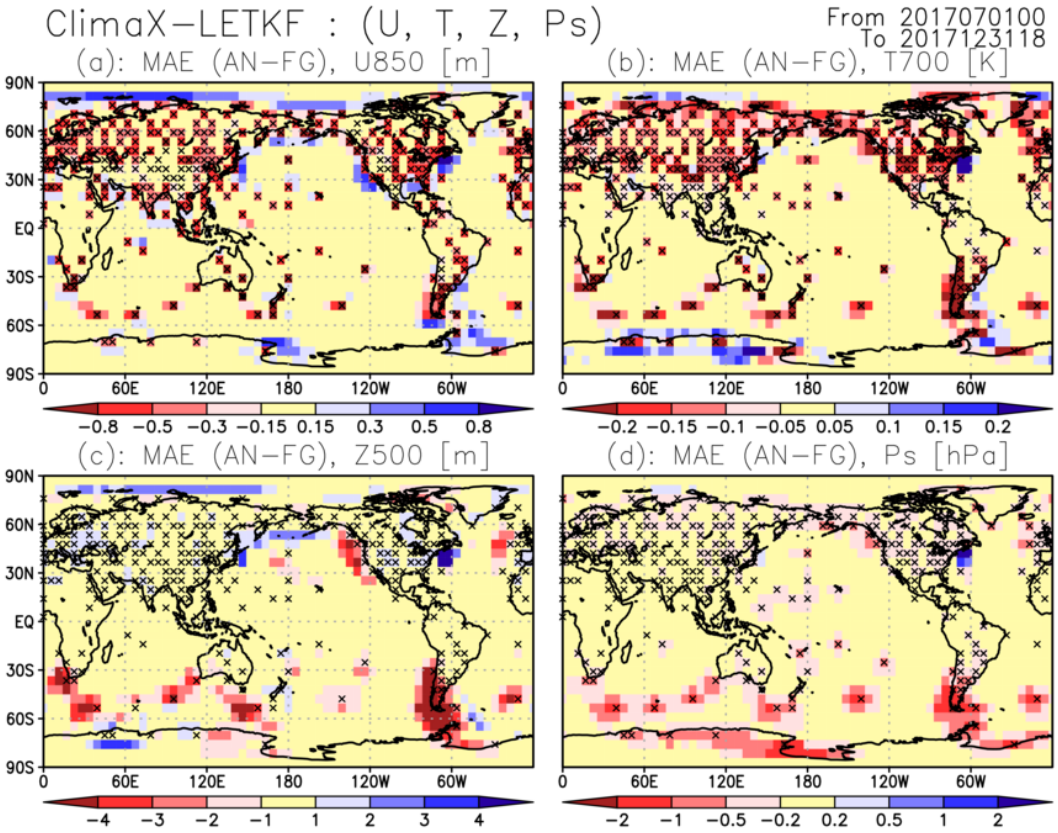


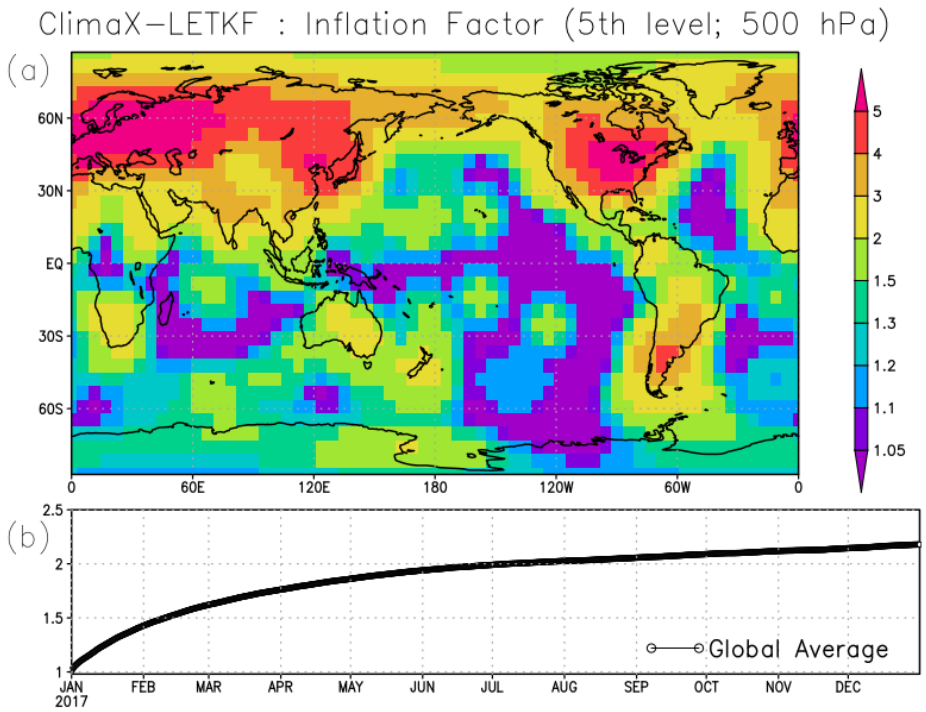
Figure 5: Global mean root mean square errors (RMSEs) verified against WeatherBench (WB) for (a) zonal wind (m/s), (b) meridional wind (m/s), (c) temperature (K), (d) specific humidity (g/kg), (e) geopotential height (m), and (f) surface pressure (hPa), as a function of the horizontal localization scales (km) averaged over July–December 2017. Colored bars and black diamonds indicate analysis (AN) and first-guess (FG) RMSEs, respectively. Blue, green, red, and purple bars in (a-e) represent 2nd, 3rd, 5th and 6th model levels (850, 700, 500, and 250 hPa, respectively). Gray bars in (f) represent surface pressure. The RMSEs of specific humidity at the 6th model level in (d) were too low to be shown.



412

413 **Figure 6:** Spatial patterns of difference between analysis (AN) and first-guess (FG) mean absolute errors (MAEs) for (a) zonal
414 wind (m/s) at 850 hPa, (b) temperature (K) at 700 hPa, (c) geopotential height (m) at 500 hPa, and surface pressure (hPa),
415 averaged over July–December 2017. Warm and cold colors represent improvements and degradations due to data assimilation.
416 Results are for a localization scale of L_h = 500 km. Black crosses indicate observing stations.

417



418

419 **Figure 7:** (a) Spatial pattern of the multiplicative inflation factor at the end of experiment on 1800UTC of December 31, 2017.

420 (b) Time series of globally averaged inflation factors. Results are for a localization scale of L_h = 600 km.