

Supplementary Information: FuXi Weather

Xiuyu Sun^{1†}, Xiaohui Zhong^{2†}, Xiaoze Xu^{3,1,4}, Yuanqing Huang¹, Hao Li^{2,1*}, J. David Neelin^{5*}, Deliang Chen^{6,7}, Jie Feng^{8,1}, Wei Han^{4*}, Libo Wu^{9,10,11,*} and Yuan Qi^{2,1*}

¹Shanghai Academy of Artificial Intelligence for Science, Shanghai, 200232, China.

²Artificial Intelligence Innovation and Incubation Institute, Fudan University, Shanghai, 200433, China.

³School of Atmospheric Physics, Nanjing University of Information Science and Technology, Nanjing, 210044, China.

⁴Earth System Modeling and Prediction Centre, China Meteorological Administration, Beijing, 100081, China.

⁵Department of Atmospheric and Oceanic Sciences, University of California, Los Angeles, 90095, USA.

⁶Ministry of Education Key Laboratory for Earth System Modeling, Department of Earth System Science, Tsinghua University, Beijing, 100084, China.

⁷University of Gothenburg, Gothenburg, Sweden.

⁸Department of Atmospheric and Oceanic Sciences and Institute of Atmospheric Sciences, Fudan University, Shanghai, 200433, China.

⁹School of Data Science, Fudan University, Shanghai, 200433, China.

¹⁰Institute for Big Data, Fudan University, Shanghai, 200433, China.

¹¹MOE Laboratory for National Development and Intelligent Governance, Fudan University, Shanghai, 200433, China.

*Corresponding author(s). E-mail(s): lihao_lh@fudan.edu.cn; neelin@atmos.ucla.edu; hanwei@cma.gov.cn; wulibo@fudan.edu.cn; qiyuan@fudan.edu.cn;

Contributing authors: xiuyu.sxy@gmail.com; x7zhong@gmail.com;
xuxiaoze1998@163.com; yuanqinghuang828@gmail.com;
deliang@gvc.gu.se; fengjiefj@fudan.edu.cn;

[†]These authors contributed equally to this work.

1 Supplementary Methods

1.1 ECMWF Reanalysis v5 (ERA5)

The fifth generation of the European Centre for Medium-Range Weather Forecasts (ECMWF) reanalysis dataset, ERA5, with a temporal resolution of 1 hour and a spatial resolution of approximately 31 km, offers data from January 1950 to the present day [1]. ERA5 is widely regarded as the most comprehensive and accurate global reanalysis archive. In this study, we used the 6-hourly ERA5 dataset at a spatial resolution of $0.25^\circ (720 \times 1440$ latitude-longitude grid points).

The FuXi model predicts a total of 70 variables, including five upper-air atmospheric variables across 13 pressure levels (50, 100, 150, 200, 250, 300, 400, 500, 600, 700, 850, 925, and 1000 hPa), and five surface variables. The upper-air atmospheric variables include geopotential (Z), temperature (T), u component of wind (U), v component of wind (V), and relative humidity (R). The surface variables include 2-meter temperature (T2M), mean sea-level pressure (MSLP), 10-meter u wind component (U10), 10-meter v wind component (V10), and total precipitation (TP). A comprehensive list of these variables and their corresponding abbreviations is given in Table 1 of the main text.

The FuXi-DA model was trained using data from June 1, 2022, to June 30, 2023. To ensure robust testing, the model utilized a comprehensive dataset covering an entire year, from July 1, 2023, to June 30, 2024. Despite the relatively limited 2-year dataset for model development, reserving one year for testing is crucial to ensure a thorough evaluation across all four seasons. In the FuXi Weather system, cycling forecasts were initialized using all-zero tensors as the initial conditions. To account for the spin-up phase of the model, common in physics-based Numerical Weather Prediction (NWP) and Data Assimilation (DA) systems, the first 2 days of cycling analysis and forecasts were excluded from the evaluation to ensure the forecasting system had reached its normal operating mode. As illustrated in Supplementary Fig. 8, the root-mean-squared-error (RMSE) values initially started high on July 1, 2023, then progressively declined, quickly reaching a state of statistical equilibrium. Consequently, the evaluation period effectively covers approximately one year, from July 03, 2023, to June 30, 2024.

1.2 Satellite data

This subsection describes all the satellite data utilized in this study. Over the past few decades, satellite observations have become increasingly important in the DA process of numerical prediction systems, primarily owing to their high spatiotemporal resolution and extensive spatial coverage. These observations have emerged as the predominant data source, markedly enhancing the accuracy of NWP forecasts [2, 3]. Joo et al. [4] demonstrated that satellite observations account for 64% of the error reduction in short-range forecasts, in contrast to 36% for surface-based observations. Notably, polar-orbiting satellites contribute 85% of the data assimilated into global NWP models [5–7].

Polar-orbiting satellites cross the equator at consistent local solar time (LST) every day [8], providing two observations per day for a given location. These satellites are categorized by their LST when crossing the equator: early morning (EM) satellites around 06:00 (HH:MM) LST, mid-morning satellites (AM) at approximately 10:00 LST, and afternoon (PM) satellites at approximately 14:00 LST [8]. Operational global DA systems require assimilating observational data every 6 hours to update the initial conditions for NWP models. A two-orbit system covers approximately 80% of the global area, leaving gaps lacking observations. Therefore, an optimal polar operational constellation should be a three-orbit system (EM, AM, and PM) [9, 10].

In this study, we selected FengYun-3E (FY-3E) as the EM satellite, Meteorological Operational Polar Satellite - C (Metop-C) as the AM satellite, and National Oceanic and Atmospheric Administration - 20 (NOAA-20) as the PM satellite. Together, the three satellites provide full global coverage in every 8-hour DA window. Microwave sounders are relatively insensitive to cloud and have substantially improved NWP forecasts, representing the most significant contributors among various satellite instruments [3, 11]. Consequently, the microwave sounders onboard these three selected satellites were used for DA in this study, as detailed in Supplementary Table 1. Figure 8 of the main text illustrates the spatial coverage of data collected by these polar-orbiting satellites, which was used to generate the analysis fields for 12 UTC on June 1, 2023. As shown in the figure, data from these satellites, gathered within an 8-hour observation window, provides almost global coverage.

In this study, we employed the brightness temperatures from the three polar-orbiting satellites for DA.

- FY-3E (EM orbit): The FY-3E, launched on 5 July 2021 from the Jiuquan Satellite Launch Centre, is the first civil meteorological satellite in EM orbit, filling a critical gap in global polar-orbiting satellite coverage [12–14]. It crosses the equator between 05:30 and 05:50 LST and carries 11 scientific instruments. For this study, we utilized all 17 channels of the Microwave Temperature Sounder-III (MWTS-III) and 15 channels of the Microwave Humidity Sounder-II (MWHs-II).

• Metop-C (AM orbit): The Metop-C launched on 7 November 2018, is the third satellite in a series of three polar-orbiting meteorological satellites developed collaboratively by the European Space Agency (ESA) and the European Organization for the Exploitation of Meteorological Satellites (EUMETSAT) as part of the EUMETSAT Polar System (EPS) [15]. Crossing the equator at approximately 09:30 LST [15], the Metop-C has eight instruments onboard. All five channels from the Microwave Humidity Sounder (MHS) and 15 channels from the Advanced Microwave Sounding Unit-A (AMSU-A) were used in this study.

• NOAA-20 (PM orbit): The NOAA-20, formerly known as JPSS-1, was launched on 18 November 2017. It is the first satellite of the U.S. next generation, polar-orbiting, environmental satellite system under the Joint Polar Satellite System (JPSS) program [16, 17]. The NOAA-20, which crosses the equator at 13:30 LST, is equipped with five instruments [18], including the Advanced Technology Microwave Sounder (ATMS). The ATMS has 22 channels: the first 16 channels are primarily used for temperature sounding from the surface to approximately 1 hPa (45 km) and the remaining channels (17–22) for humidity sounding in the troposphere from the surface to approximately 200 hPa (10 km) [5]. This provides high-resolution measurements of temperature and moisture and offers detailed insights into tropical cyclone (TC) warm cores and rainfall intensity when measuring conditions within the eye of a TC.

Radio occultation (RO) is a cost-effective remote-sensing technique that accurately measures the atmospheric gradient of atmospheric refractivity. This measurement is essential for deriving vertical profiles of temperature, pressure, and humidity [19–21]. In this technique, signals transmitted by a Global Navigation Satellite System (GNSS) satellite and received by a low-Earth orbiting satellite, are refracted by the Earth’s atmosphere. This refraction alters the path and timing of the signals. The precise positions of both the transmitting and receiving satellites enable the accurate measurement of signal delays and bending angles, which are crucial to derive weather parameters such as temperature, pressure, and humidity. RO data offers global three-dimensional (3D) coverage, high accuracy, and high vertical resolution. Unlike most remote-sensing measurements, which are often negatively affected by surface weather conditions such as clouds and precipitation, RO measurements maintain their integrity under all weather conditions, ensuring continuous data acquisition. Global Navigation Satellite System RO data produce high-resolution profiles from the surface of the Earth up to the stratopause, an atmospheric layer typically characterized by sparse radiance data and poor NWP model performance, making them highly complementary to other sounding data (e.g., data from microwave sounders) within operational DA systems [3, 22, 23]. In this study, we used vertical profiles of refractivity retrieved from GNSS-RO. However, GNSS-RO data are inhomogeneous, non-uniform, and sparsely distributed globally, appearing as discrete data points across various atmospheric

Supplementary Table 1: A summary of polar-orbiting satellite information.

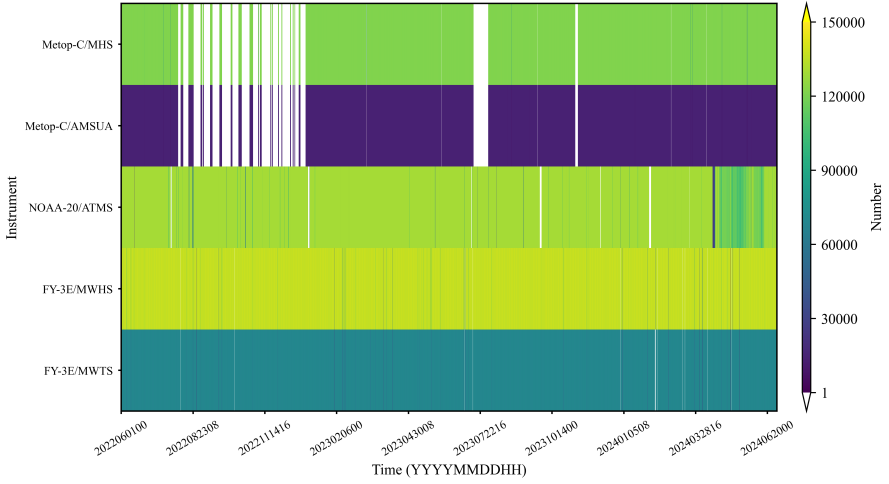
Satellite	Orbit	Instruments	Channels	Preprocessed Data Dimensions
FY-3E	Polar-orbiting (early morning)	MWTS-III	17	$8 \times 20 \times 720 \times 1440$
		MWHS-II	15	$8 \times 18 \times 720 \times 1440$
Metop-C	Polar-orbiting (morning)	AMSU-A	15	$8 \times 18 \times 720 \times 1440$
		MHS	5	$8 \times 8 \times 720 \times 1440$
NOAA-20	Polar-orbiting (afternoon)	ATMS	22	$8 \times 25 \times 720 \times 1440$
GNSS-RO	-	-	512	$8 \times 32 \times 720 \times 1440$

columns [24]. Therefore, specialized preprocessing techniques are required to transform GNSS-RO data into a gridded format, which simplifies integration into the FuXi-DA model, and aligns with the assimilation of other satellite data. This study utilized RO data collected from multiple GNSS satellite receivers. The availability timeline, indicating periods when data were either available or missing for each receiver, is illustrated in Supplementary Fig. 2.

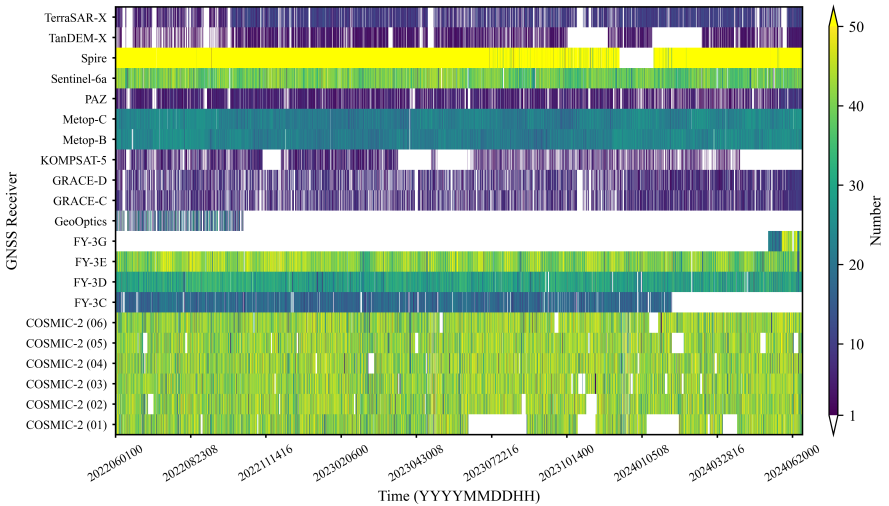
1.3 Satellite data availability

Supplementary Figs. 1 and 2 present timelines of satellite data availability of the whole dataset for the development of FuXi Weather, covering the period June 1, 2022 - June 30, 2024. Among the three polar-orbiting satellites, the Metop-C satellite exhibited greater data unavailability compared with the other two satellites, especially during the training phase.

Supplementary Fig. 2 illustrates the data availability from various meteorological satellites, including several commercial constellations. The Constellation Observing System for Meteorology, Ionosphere, and Climate 2 (COSMIC-2) [25, 26], a joint mission by Chinese Taiwan and the U.S., is one such commercial satellite constellation. Launched on 25 June 2019, COSMIC-2 includes six satellites to collect RO data using GNSS signals. This constellation notably contributes to global weather prediction, ionospheric research, and climate monitoring efforts.



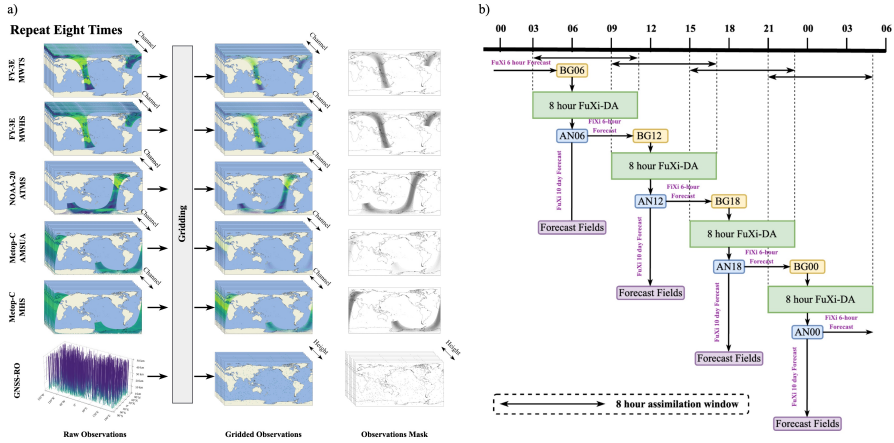
Supplementary Figure 1: Satellite data availability from three polar-orbiting meteorological satellites. Data availability for FengYun-3E, Meteorological Operational Polar Satellite-C (Metop-C), and National Oceanic and Atmospheric Administration-20 (NOAA-20) for the 1-year testing period. The colors denote the number of observations at the temporal resolution of 1 hour.



Supplementary Figure 2: Satellite data availability for the 21 Global Navigation Satellite System (GNSS) receivers used in the 1-year testing period. The colors denote the number of observations at the temporal resolution of 1 hour.

2 Supplementary FuXi Weather

Supplementary Fig. 3 illustrates FuXi Weather, designed for generating 6-hourly updated global weather forecasts. This system includes three main components: satellite data preprocessing, DA via the FuXi-DA model, and weather forecasting via the FuXi model. During the DA phase, the FuXi-DA model combines preprocessed satellite data with prior (background) forecasts to produce analysis fields. These fields are then utilized by the FuXi model to generate 10-day weather forecasts.



Supplementary Figure 3: Overall structure of the FuXi Weather system. (a) Satellite data preprocessing, incorporating data from three polar-orbiting meteorological satellites (FY-3E, Metop-C, and NOAA-20), and GNSS-RO data. (b) Data assimilation and forecasting cycles with the FuXi-DA model producing analysis fields, followed by forecast generation via the FuXi model.

2.1 Satellite data preprocessing

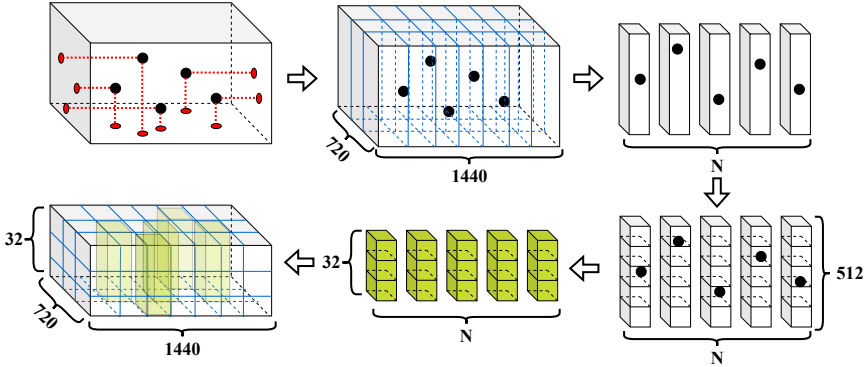
This subsection elaborates on the data preprocessing process applied to satellite data described in Section 1.2. Figure 8 of the main text shows the spatial coverage of FY-3E, Metop-C, and NOAA-20, denoted by blue, red, and green colors, respectively, over 8 hours from 3 hours before to 4 hours after 12 UTC on June 1, 2023. As shown in Figure 8 of the main text, satellite data are heterogeneous in space and time; hence, nearest-neighbor interpolation from satellite data to the FuXi output grids must be performed. Data from the three polar-orbiting satellites were averaged within grid points corresponding to the FuXi forecasts (720×1440 in latitude and longitude) to represent the satellite observations at each location. All data collected by an instrument onboard a specific satellite within this 8-hour assimilation window were concatenated before being input to FuXi-DA.

Preprocessing GNSS-RO data poses considerable challenges owing to its inhomogeneous, non-uniform, and sparse global distribution. Global Navigation Satellite System RO data are discrete points [27], defined by latitude, longitude, and altitude, structurally resembling 3D point clouds generated by 3D sensors such as light detection and ranging (LiDAR)-based devices, as shown in Figure 8 of the main text. A point cloud is a collection of data points in 3D space, typically represented by Cartesian coordinates along the x, y, and z axes, and potentially includes additional features such as surface normals, red-green-blue values, and timestamps [28, 29]. The adoption of this format has surged in popularity owing to the growing availability and expanding applications of 3D sensing technologies in fields such as robotics, autonomous driving, and augmented/virtual reality [30]. Machine learning, which has been successful with two-dimensional (2D) data, faces considerable challenges when applied to 3D point clouds such as GNSS-RO data, which are characterized by an irregular global distribution and unstructured nature.

In this study, we adapted the PointPillars method [31] to process the refractivity profiles retrieved from GNSS-RO data. PointPillars organizes sparse point cloud data into vertical columns, referred to as "pillars", by discretizing the horizontal plane into a grid and aggregating points within each grid cell into compact feature representations. These features are then transformed into a 2D pseudo-image format. This transformation allows 2D convolutional neural networks (CNNs) to extract high-level spatial features from otherwise irregular and sparse input data. As illustrated in Supplementary Fig. 4, we applied this approach to GNSS-RO data by first discretizing the refractivity profiles and aligning them with the $0.25^\circ \times 0.25^\circ$ grid (720×1440 latitude-longitude points) used in the FuXi model. The data were then organized into vertical pillars, each representing a vertical column of refractivity measurements at a specific grid point. Due to the sparse distribution of GNSS-RO data, most grid columns contained no data, with only a small fraction containing data. This sparsity is visualized in Figure 8 of the main text, where yellow dots indicate the presence of GNSS-RO observations within the 8-hour DA window, spanning from 3 hours before to 4 hours after 12 UTC on June 1, 2023. Notably, only 0.59% ($\frac{6902}{720 \times 1440}$) of the grid columns contained GNSS-RO data.

Given the high vertical resolution of GNSS-RO profiles (approximately 100 m[20]), which greatly exceeds that of the FuXi background fields, the refractivity profiles were interpolated to uniform height levels up to 50 km altitude. This preprocessing step produced a 2D tensor of dimensions $512 \times N$, where 512 and N correspond to the number of vertical layers and the number of samples, respectively. To extract informative features from these vertical profiles, we employed a fully connected layer followed by layer normalization [32] to each pillar's data, yielding a tensor of reduced dimensions $32 \times N$, where 32 denotes the number of encoded feature channels. These encoded features were then mapped back to the FuXi model's 2D grid, yielding a tensor of dimensions $32 \times 720 \times 1440$, with 720 and 1440 representing the latitude and longitude

241 grid points, respectively. This transformation yielded a high-level representa-
 242 tion of the GNSS-RO data aligned with the spatial resolution of the FuXi
 243 background forecasts. By leveraging the PointPillars framework, we efficiently
 244 addressed the sparsity and irregularity of GNSS-RO data, facilitating seamless
 245 integration into the FuXi-DA model for subsequent processing.



Supplementary Figure 4: Data preprocessing for Global Navigation Satellite System (GNSS) radio occultation (RO) data through the PointPillars method.

246 To mitigate observational gaps caused by unscanned areas or data loss,
 247 we implemented a masking technique in our satellite data preprocessing. This
 248 technique assigns a value of 1 to grid points having available data and 0 where
 249 data are missing; this is crucial for the effective assimilation of satellite obser-
 250 vations into the FuXi-DA model. These preprocessing steps are schematically
 251 illustrated in Supplementary Fig. 3a. The polar-orbiting satellite data were
 252 normalized using the z-score normalization technique, ensuring uniformity in
 253 their mean and variance. The GNSS-RO refractivity data were normalized by
 254 dividing the raw data by 360.

255 It is important to note that, aside from excluding satellite-observed bright-
 256 ness temperatures above 350 K and below 50 K, no additional quality control
 257 screening is applied to the observational data input to FuXi Weather. Conse-
 258 quently, FuXi Weather operates as an all-pixel, all-surface, all-channel, all-sky
 259 machine learning DA and weather forecasting system.

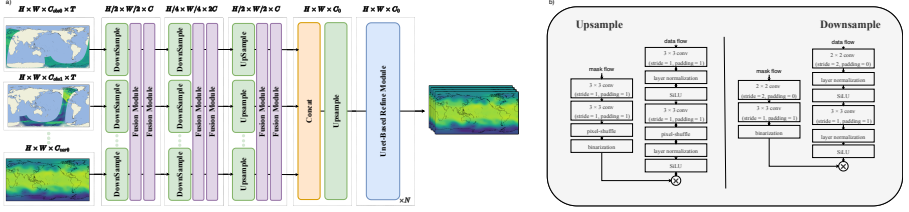
260 2.2 FuXi-DA model architecture

261 Similar to conventional DA methods, the FuXi-DA model uses both back-
 262 ground forecasts and observations as inputs. The background forecasts,
 263 generated by the FuXi model [33], encompass both upper-air and surface
 264 variables arranged in a data cube with dimensions of $70 \times 720 \times 1440$ (vari-
 265 ables, latitude, longitude). Unlike the single data cube used by FuXi, FuXi-DA

employs six separate data cubes: a $5 \times 720 \times 1440$ cube for five surface variables, and five $13 \times 720 \times 1440$ cubes for 5 upper-air atmospheric variables (Z, T, U), V, and R) across 13 pressure levels. Satellite observations from six sources, including five instruments on the three-orbit system and GNSS-RO data, are structured into tensors reflecting time frames, observation channels, and geographic coordinates (latitude and longitude). For example, the tensor dimensions for MWTS-III on FY-3E are $8 \times 20 \times 720 \times 1440$, covering 17 microwave channels and three types of encoded observational information (latitude, longitude, and satellite scan angle). The preprocessed data dimensions of all satellite data used in this study are summarized in Supplementary Table 1. The encoded observational information is included because biases in satellite data often vary with scan angle and geographic location [34, 35]. The time frames are consistently set at eight to align with the 8-hour assimilation window of FuXi-DA. Before integration into the FuXi-DA model, background forecast dimensions are adjusted from 721×1440 to 720×1440 using bilinear interpolation to ensure consistency with the observational data. Before input to the FuXi-DA model, the temporal and observational information of satellite data are merged. For example, data dimensions for MWTS-III on FY-3E are restructured from $8 \times 20 \times 720 \times 1440$ to $160 \times 720 \times 1440$.

In traditional DA frameworks, the assimilation of satellite radiance requires observational operators to map background forecasts from the state vector to observational space. In contrast, FuXi-DA eliminates the need for conventional observational operators by converting observations and forecasts into separate latent spaces and merging their information using fusion modules. The FuXi-DA model employs a multi-branch architecture, as illustrated in Supplementary Fig. 5b, incorporating dedicated branches for observations and background forecasts. Specifically, it includes six observation branches, each designed to process a set of six satellite observations: five from microwave sounders on a three-orbit system and one from GNSS-RO data. An additional six branches handle background forecasts. This architecture is scalable, allowing for the integration of additional satellite data not included in this study by adding more observation branches. As illustrated in Supplementary Fig. 5, each branch processes data using two repeated blocks consisting of a downsampling layer followed by two fusion modules, then an up-sampling layer with two additional fusion modules. The features from all branches are then concatenated and passed through a further up-sampling layer. Finally, a refinement module, composed of three repeated U-net blocks, further improves the accuracy of the analysis fields. Interactions among features across branches are optimized through ‘bottleneck fusion’ [36] across multiple layers, enhancing the integration of multi-modal data.

Specifically, fusion modules adjust the weighting of observations and background data across various scales using stage-by-stage corrected condition information. As shown in Supplementary Fig. 3b, these modules leverage the U-net’s down-sampling and up-sampling capabilities of the U-net architecture



Supplementary Figure 5: Structure of the FuXi-DA model. (a) Overall model architecture. (b) Detailed designs of up-sampling and down-sampling layers.

to enhance information interaction between branches and improve feature resolution through skip connections. A fusion module generates two outputs: one containing analysis increments added to the background branch and the other consisting of observed bias information added to the observation branch. This operation is based on the traditional bias correction process, in which predictions computed from the background field are typically used for bias correction [34, 35, 37].

In this study, the fusion module comprised one U-Net block [38], with two down-sampling and two up-sampling layers connected by skip connections to preserve local information. Each down-sampling layer includes a sequence of operations: a 2D convolution layer with a kernel size of 2 and stride of 2, followed by a layer normalization layer [32], then a Sigmoid-weighted Linear Unit (SiLU) activation [39], and another 2D convolution layer with a 3×3 kernel and stride of 1. Data dimensions are sequentially reduced from $70 \times 1440 \times 720$ to $256 \times 720 \times 360$, and further to $512 \times 360 \times 180$. Each up-sampling layer consists of a 2D convolution layer with a 3×3 kernel and stride of 1, layer normalization, SiLU activation, another 2D convolution layer (kernel size 3×3 , stride 1), and a pixel-shuffle layer [40] with an upscaling factor of 2. This pixel-shuffle layer rearranges elements in a tensor of shape $(B, C \times r^2, H, W)$ into a tensor of shape $(B, C, H \times r, W \times r)$, where B and r denote the batch size and upscaling factor, respectively. Subsequently, the data dimensions after two upsampling layers are restored to $256 \times 720 \times 360$ and finally to $70 \times 1440 \times 720$. In total, the FuXi-DA model consists of 0.7 billion parameters.

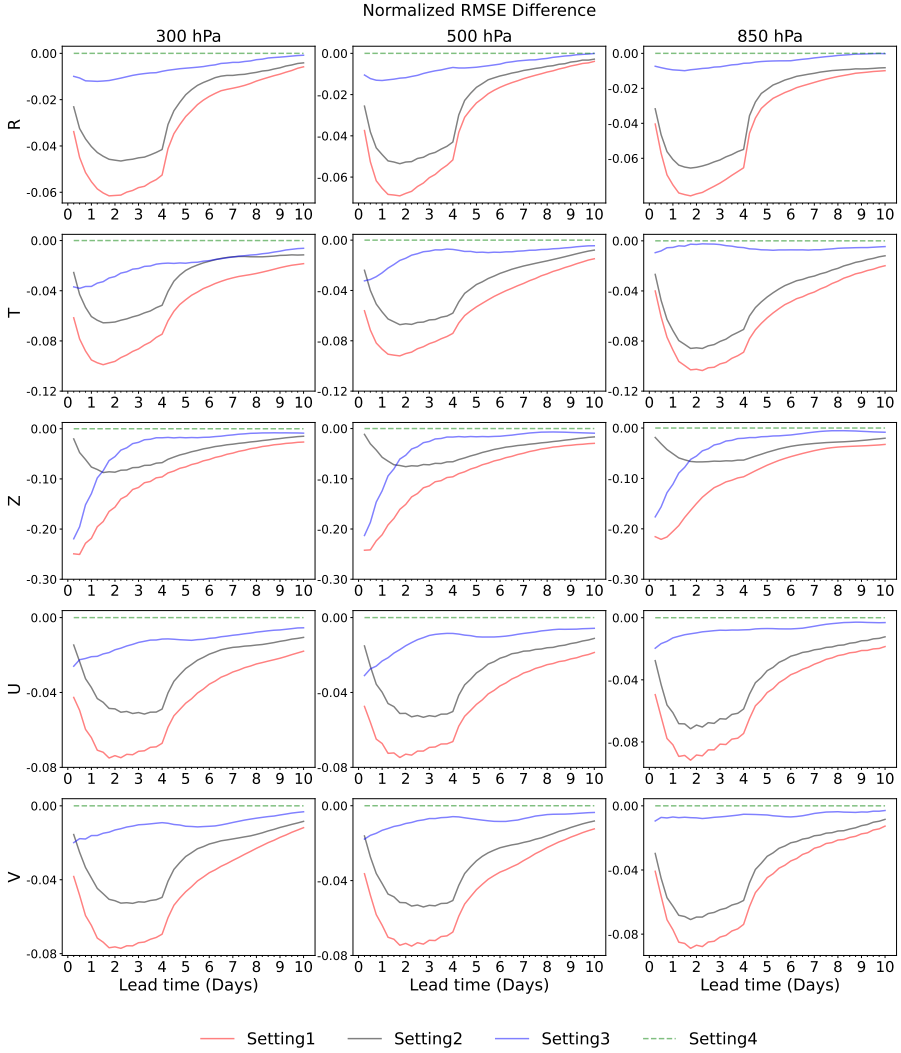
The FuXi-DA model performs global DA four times per day at 00, 06, 12, and 18 UTC, each within an 8-hour assimilation window that starts 3 hours before and ends 4 hours after the forecast initialization time, with each observation representing conditions over a one-hour interval centered on the observation time. For example, the initial conditions at 12 UTC incorporate all satellite observations from 08:30 to 16:30 UTC. This rigorous schedule ensures a comprehensive global analysis at 0.25° resolution for the FuXi forecast model, effectively incorporating all relevant satellite observations and adjusting for known biases related to scan angle and geographic location. The selection of an 8 hour DA window leads to some observations being used more than once during the assimilation process. This sub-daily interval was selected

primarily for practical considerations related to the implementation of deep learning algorithms. The number eight was chosen because it is a power of two, which aligns with the architectural design of graphics processing units (GPUs). Dimensions that are powers of two enable more efficient memory utilization and computational performance, owing to the structure of parallel processing on GPUs.

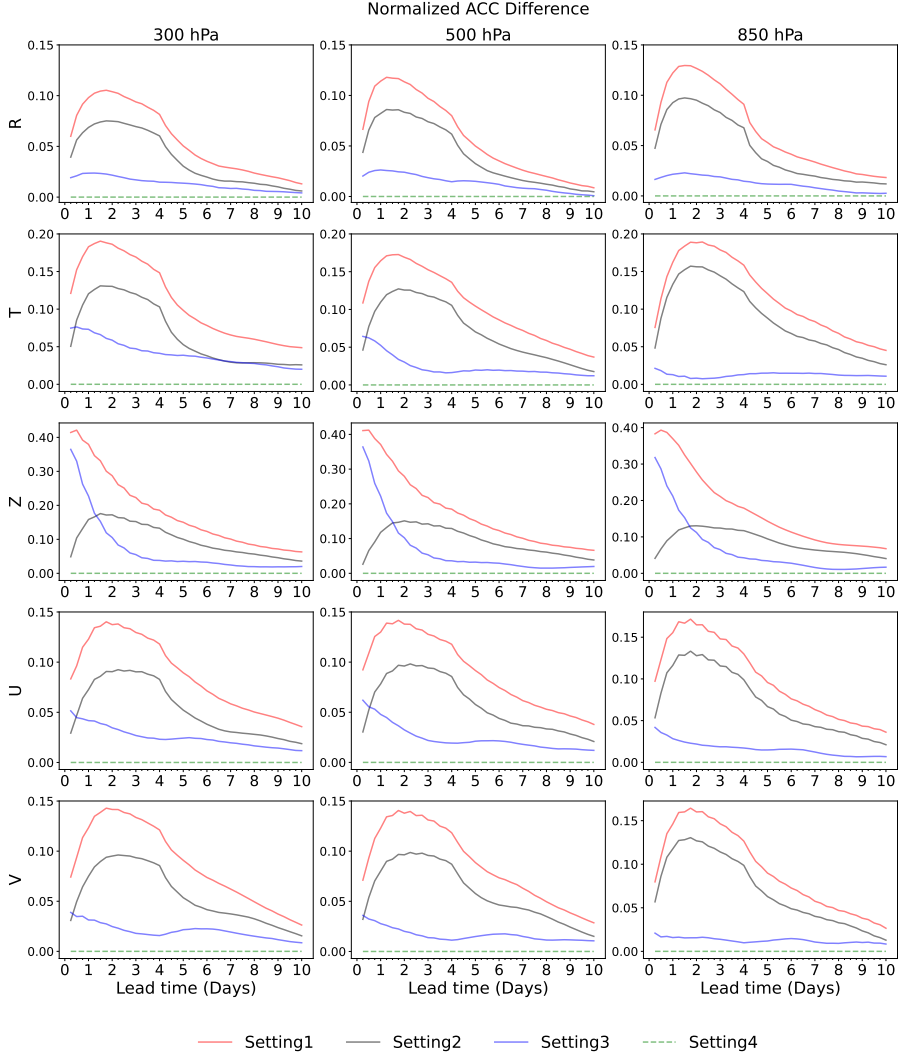
Training FuXi-DA with a non-overlapping 6-hour DA window and evaluating its performance (not shown) over one month produces RMSE values comparable to those obtained using an 8-hour DA window. This suggests that machine learning-based DA’s performance is less sensitive to the DA window overlap.

2.3 FuXi model

Following the generation of analysis fields by FuXi-DA, the FuXi model is employed to produce 10-day forecasts. As described by Chen et al. [33], FuXi consists of three cascade models—FuXi-Short, FuXi-Medium, and FuXi-Long—optimized for specific forecast time windows of 0–5, 5–10, and 10–15 days, respectively. Specifically, when ERA5 serves as the initial conditions, the 5-day forecasts from FuXi-Short are utilized as the starting point for FuXi-Medium, which forecasts the subsequent 5–10 days. However, the analysis fields from FuXi-DA are less accurate than ERA5, resulting in a marked degradation in forecast performance. This discrepancy is particularly evident in the reduced accuracy of the 5-day forecasts from FuXi-Short when based on FuXi-DA analysis fields (see main text Fig. 2). To mitigate this issue, the FuXi-Short model is fine-tuned using the FuXi-DA analysis fields (details in Section 3.3 below). Moreover, we found that 4-day forecasts from FuXi-Short, when initialized with FuXi-DA analysis fields, achieved comparable accuracy to the 5-day forecasts initialized with ERA5. Consequently, we employed these 4-day forecasts as the initial conditions for FuXi-Medium for 4–10-day predictions. The impact of this adjustment is shown in Supplementary Figs. 6 and 7, where it can be seen that using FuXi-Short for 4-day forecasts leads to improved accuracy in 4–10-day predictions, reflected in lower RMSE and higher anomaly correlation coefficient (ACC) values compared with those when using FuXi-Short for 5-day forecasts. Although further fine-tuning of FuXi-Medium could theoretically further enhance performance, this is beyond the scope of this paper.



Supplementary Figure 6: Comparison of forecast performance between the use of FuXi-Short for 4-day forecasts and the use of FuXi-Short for 5-day forecasts. The figure presents the normalized differences in the globally-averaged and latitude-weighted root-mean-square error (RMSE). The comparison spans a 1-year testing period (July 03, 2023–June 30, 2024). The analysis includes five variables: relative humidity (R), temperature (T), geopotential (Z), u component of wind (U), and v component of wind (V), at three pressure levels (300, 500, and 850 hPa). The five rows and three columns correspond to the five variables and three pressure levels, respectively. FuXi forecasts (red and gray lines) are verified against ERA5. The performance change on days 4 and 5 reflects the transition from FuXi-Short to FuXi-Medium for the respective forecasts (red and gray lines).



Supplementary Figure 7: Comparison of forecast performance between the use of FuXi-Short for 4-day forecasts and the use of FuXi-Short for 5-day forecasts. The figure presents the normalized differences in the globally-averaged and latitude-weighted anomaly correlation coefficient (ACC). The comparison spans a 1-year testing period (July 03, 2023-June 30, 2024). The analysis includes five variables: relative humidity (RH), temperature (T), geopotential (Z), u component of wind (U), and v component of wind (V), at three pressure levels (300 hPa, 500 hPa, and 850 hPa). The five rows and three columns correspond to five variables and three pressure levels, respectively. The performance change on days 4 and 5 reflects the transition from FuXi-Short to FuXi-Medium for the respective forecasts (red and gray lines).

3 FuXi Weather training

3.1 FuXi-DA model training

As discussed previously, FuXi-DA operates on a 6-hour cycle, utilizing the prior 6-hour forecasts generated by FuXi Weather. The simplest approach to train the FuXi-DA model involves using these 6-hour forecasts, initialized with ERA5, as input. However, the effectiveness of this approach relies on the assumption that FuXi forecasts initialized with ERA5 are comparable in accuracy to those initialized with the analysis fields produced by FuXi-DA. It is important to note that the FuXi-DA model assimilates considerably fewer observational data compared with those used to generate ERA5, potentially resulting in discrepancies in accuracy. To better simulate operational scenarios in which FuXi-DA is used during the 6-hourly DA and forecasting cycles, we randomly sampled forecasts from the FuXi model (initialized with ERA5) over lead times ranging from 6 hours to 5 days and used these sampled forecasts as background forecasts to train FuXi-DA. We also initialized the model training with all-zero tensors for state variables, setting all variables to zero at the start of the cycling DA and forecasting processes for FuXi Weather.

Training of the FuXi-DA model was performed in two stages on four Nvidia A100 GPUs. During the first stage, the model underwent 24,000 gradient descent updates, with each GPU processing a batch size of 1. The model was developed using the PyTorch framework [41], employing the AdamW optimizer with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay coefficient of 10^{-5} . The learning rate followed a warm-up and cosine annealing schedule [42], starting at 10^{-8} , increasing linearly to peak at 2×10^{-3} after 500 steps, and then decreasing according to a cosine annealing schedule over 24,000 iterations, totaling 48 hours. Post-training analysis revealed improved accuracy in the fields generated by the FuXi-DA model. Consequently, we adjusted our sampling strategy, reducing the maximum lead time from 5 to 3 days for the FuXi model forecasts initialized with ERA5, and excluding all-zero tensors. In the second stage, the model underwent an additional 4000 gradient descent updates over approximately 8 hours.

To calculate the discrepancies between the outputs of the FuXi-DA model and ERA5, a latitude-weighted L1 loss function was utilized, defined as follows:

$$L1 = \frac{1}{C \times H \times W} \sum_{c=1}^C \sum_{i=1}^H \sum_{j=1}^W \alpha_i |\hat{\mathbf{X}}_{c,i,j} - \mathbf{X}_{c,i,j}| \quad (1)$$

where C , H , and W denote the number of channels, latitude, and longitude of the grid points, respectively. $\hat{\mathbf{X}}$ represents the ERA5 reanalysis data, and

$\alpha_i = H \times \cos\Phi_i / \sum_{i=1}^H \cos\Phi_i$ is the latitude-specific weighting factor at latitude Φ_i .

To jointly optimize the analysis and forecasts, we implemented a loss function similar to those used in traditional 4DVar approaches. Here, the forecasting model, FuXi, provides gradients only for the DA component while keeping its parameters fixed. The final loss function used for joint optimization in the FuXi-DA and FuXi model is given by:

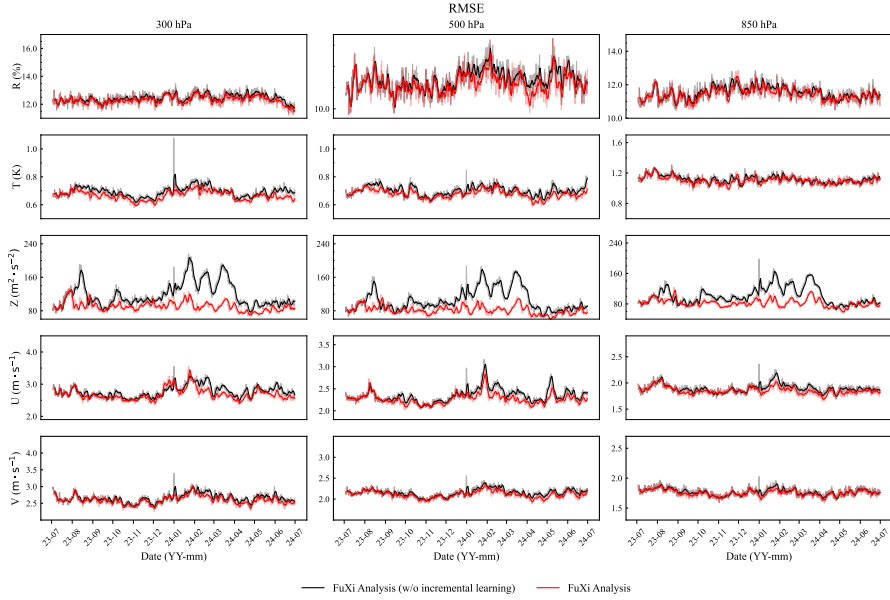
$$L = L1^0 + \frac{1}{T} \sum_{t=1}^T L1^t \quad (2)$$

where T denotes the number of time steps. When $t = 0$, $\mathbf{X}$ corresponds to the analysis field in $L1$, and when $t > 0$, it denotes the forecast field in $L1$. In this study, we set $T = 1$ to include both the analysis field and one subsequent forecast field at a 6-hour lead time. This setup incorporates both the current analysis field and one subsequent forecast field, thereby mirroring the traditional 4DVar assimilation and facilitating end-to-end training for FuXi Weather.

3.2 Incremental learning

The FuXi-DA model was developed using a 2-year dataset, with 1 year for training and 1 year for testing. This dataset is considerably smaller than the 37-year dataset used for training cascaded FuXi forecasting models [33]. Training machine learning models on such limited datasets poses substantial challenges in maintaining model robustness and accuracy, a scenario typical of few-shot learning [43, 44]. Furthermore, the quality and quantity of satellite measurements often fluctuate over time owing to changes in satellite instrument characteristics and increases in the numbers of GNSS receivers (see Supplementary Fig. 2).

These variations mean that the FuXi-DA model must learn incrementally from new data and retain previously acquired knowledge, a process known as incremental or continual learning. Incremental learning methods are generally categorized into three types: replay-based, regularization-based, and parameter isolation methods [45]. For the FuXi-DA model, we adopted a replay-based incremental learning strategy, wherein the model is retrained monthly using data from the preceding year. This includes replaying data from the previous 11 months alongside new data from the current month. For example, the December 2023 update of the FuXi-DA model, intended for use in January 2024, was trained on data spanning December 25, 2022–December 25, 2023. This training regime took approximately 8 hours for 4000 gradient descent updates. The effectiveness of this strategy is illustrated in Supplementary Fig. 8, which compares the analysis fields produced by FuXi Weather with and without incremental learning. The results show a substantial improvement in performance for FuXi Weather when employing the replay-based incremental learning strategy.



Supplementary Figure 8: Comparison of analysis fields produced by FuXi Weather with and without applying an incremental learning strategy over a 1-year testing period, spanning July 1, 2023–June 30, 2024. The figure shows the time series of the globally-averaged and latitude-weighted root-mean-square error (RMSE) for analysis fields generated by FuXi-DA with (solid red lines) and without (solid black lines) applying the incremental learning strategy. This comparison includes five variables: relative humidity (R), temperature (T), geopotential (Z), u component of wind (U), and v component of wind (V), at three pressure levels (300, 500, and 850 hPa). The five rows and three columns correspond to the five variables and three pressure levels, respectively. To improve clarity, the original data are shown with reduced opacity, while solid lines represent smoothed values using a 12-point moving average.

3.3 FuXi-Short model fine-tuning

As previously mentioned, the quality of the analysis fields generated by FuXi-DA is inferior to that of ERA5. The degree of quality degradation varies markedly across different variables, as illustrated in the main text Fig. 2, with the variable R exhibiting the highest accuracy. Consequently, we used the analysis fields produced by FuXi-DA as the initial conditions for fine-tuning the FuXi model. This fine-tuning process, following the methods used in GraphCast [46] and FuXi [33], employs an autoregressive training regime and a curriculum training schedule. The training progressively increases the number of autoregressive steps from 2 to 12, with 500 gradient descent updates at each step. Throughout this process, the model is trained at a constant learning rate of 1×10^{-7} , requiring approximately 24 hours to fine-tune on a cluster

of four Nvidia A100 GPUs. A detailed discussion of the impact of fine-tuning the FuXi-Short model is presented in Section 10.

4 Single observation tests

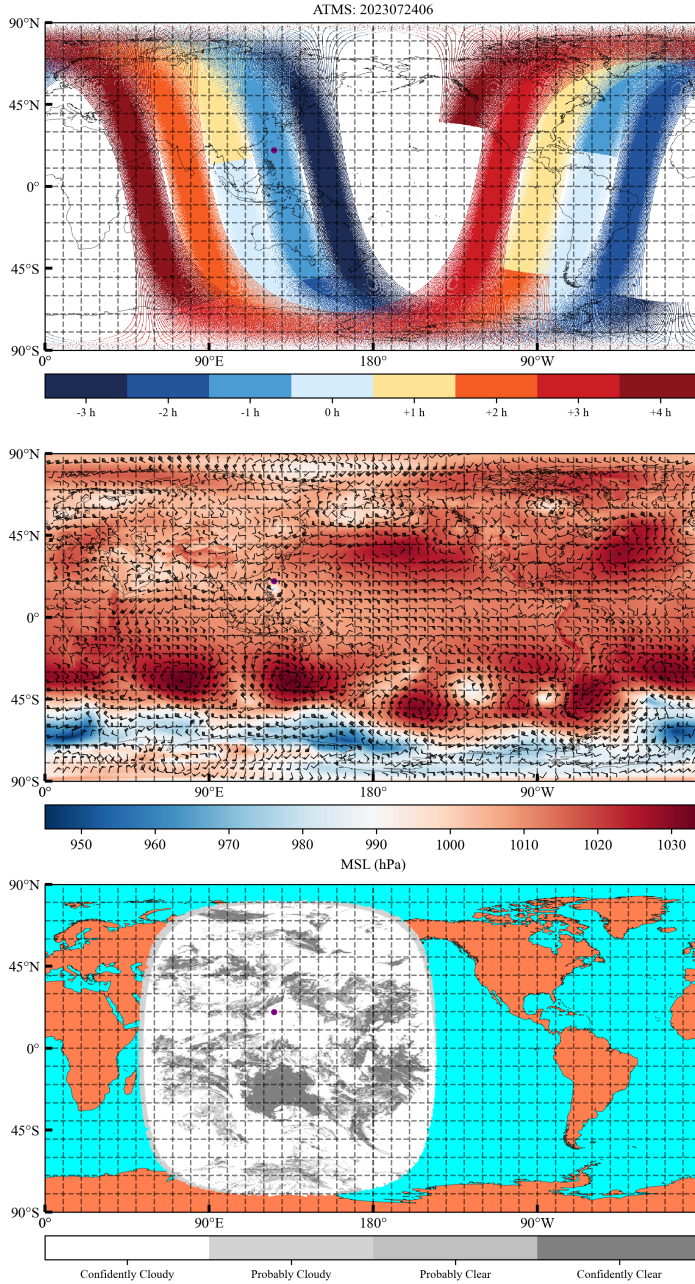
4.1 Single observation test methodology

The single observation test is a critical experiment in DA that assesses how the assimilation of one observational data point impacts the analysis field. This involves assimilating a single observation at a specific location with a predefined innovation (observation minus background) and observation error std. This test helps to evaluate the DA system’s sensitivity to specific observations, diagnose its performance, and improve our understanding of error propagation, ultimately refining the DA system. Changes in the analysis fields reveal valuable information, such as the ratio of background to observation error variance, variable correlations, and spatial structure of the background error covariance. The spatial spread of the analysis increments is largely influenced by the spatial correlations within the background error. As an initial diagnostic, the single observation test is crucial in evaluating new DA systems.

For example, in the three-dimensional variational (3D-Var) DA method, the analysis increment is given by:

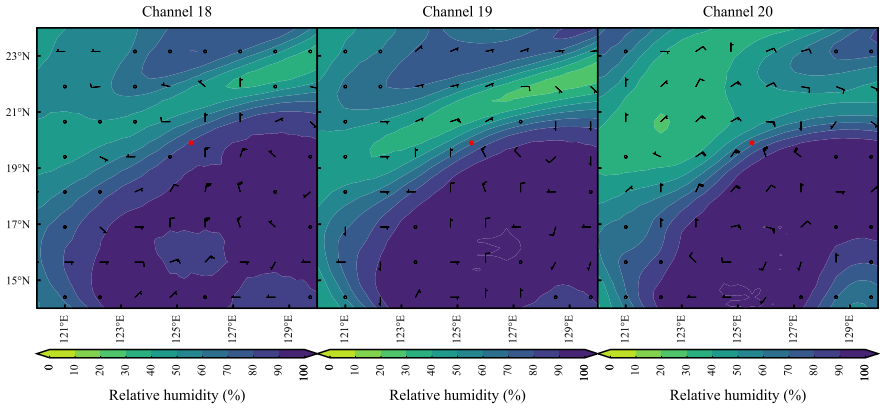
$$\mathbf{X}^a - \mathbf{X}^b = \mathbf{B}\mathbf{H}^T(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1}(\mathbf{Y}^o - H[\mathbf{X}^b]) \quad (3)$$

where, $\mathbf{X}^a$ and $\mathbf{X}^b$ represent the analysis and background fields, respectively, and $\mathbf{Y}^o$ is the observation vector. The matrices $\mathbf{B}$ and $\mathbf{R}$ denote the background and observation error covariances, respectively, while H is the observation operator that maps the model state space to the observation space. The Jacobian matrix, $\mathbf{H} = \frac{\partial H(\mathbf{x})}{\partial \mathbf{x}}$, reflects the sensitivity of the observation to the background state variables. In the case of a single observation, the term $(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1}(\mathbf{y}^o - H[\mathbf{x}^b])$ simplifies to a scalar, with $\mathbf{B}$ determining the spatial distribution of the analysis increments. For satellite radiance observations, the innovation, $\mathbf{y}^o - H[\mathbf{x}^b]$ is typically computed directly, where $\mathbf{H}$ is obtained from a fast radiative transfer model (RTM) such as the radiative transfer model for television infrared observation satellite operational vertical sounder (RTTOV) [47]. Comparing the resulting analysis increments to theoretical expectations helps to validate the performance of the DA system. In contrast, the FuXi-DA system does not explicitly incorporate background and observation error covariance, nor does it include an observation operator to convert the 13-level background fields to satellite observations, meaning that a precise innovation value cannot be defined for single observations.



Supplementary Figure 9: Single observation tests with 5 K perturbation. The perturbation is located over the ocean near Typhoon Doksuri, at 19.9° N, 125.5° E (purple dot). The first panel represents the brightness temperature observed by NOAA-20 ATMS over 8 hours, spanning from 3 hours before to 4 hours after 06 UTC on July 24, 2023. The second panel illustrates the spatial distribution of mean sea-level pressure (MSLP) with 850 hPa wind vectors overlaid. The third panel displays the cloud mask product retrieved from the Advanced Geosynchronous Radiation Imager (AGRI) aboard Fengyun-4B. The perturbation is located in a cloudy region.

To evaluate the impact of a single observation in FuXi-DA, we followed the approach proposed by Xu et al. [48], involving two runs: one with original observations and one with a small perturbation. The difference between these two analysis fields represents the analysis changes caused by the perturbation. In this study, we introduced a 5 K perturbation to NOAA-20 ATMS over the ocean near Typhoon Doksuri at 19.9°N, 125.5°E, at 05 UTC on July 24, 2023. The background field was a 6-hour forecast initialized at 00:00 UTC on July 24, 2023, following analysis cycles with the full observation set. Notably, the perturbation was applied across a 5×5 grid point rather than a single grid point. This design accounts for the strong spatial correlation of observation errors in the unthinned data. Since neighboring observations exhibit high error covariance, localized single-point perturbations would produce insufficiently representative impacts, whereas area perturbations better capture realistic error structures.

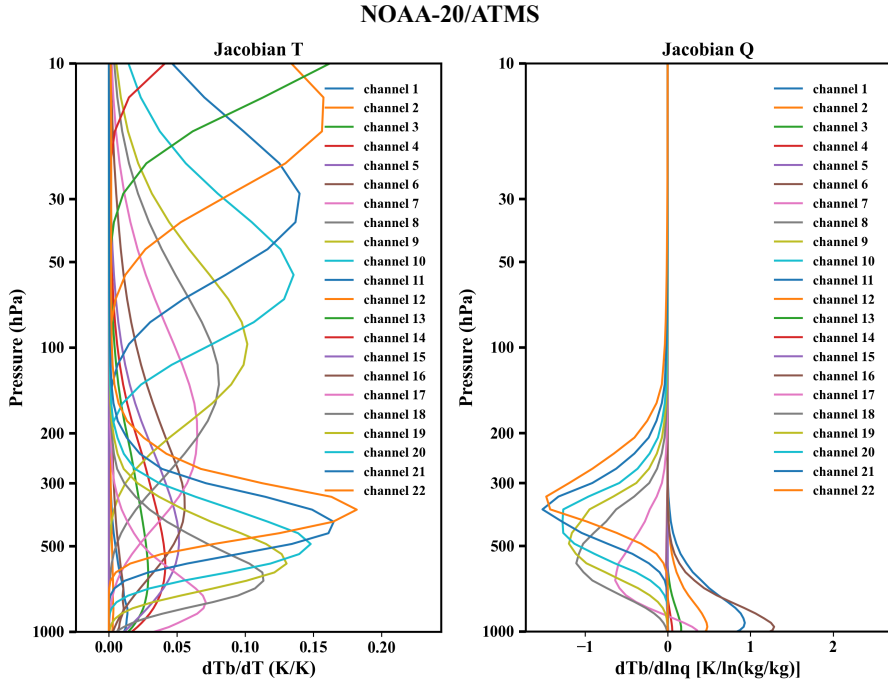


Supplementary Figure 10: Changes of wind vectors in the analysis fields resulting from a 5 K perturbation to the NOAA-20 ATMS observation at a selected location, based on the background field for 06 UTC on July 24, 2023. The figure shows the horizontal spatial distribution of wind vector changes in the analysis for channels 18 to 20 at 600, 500, and 400 hPa, overlaid with relative humidity fields. In the wind vector plots, a long barb represents 0.03 m/s, a short barb 0.015 m/s, and a pennant indicates 0.15 m/s.

4.2 Jacobians

In this study, Jacobians were derived using the RTTOV version 12.2, based on the US standard atmosphere profile; Jacobians can vary markedly across different atmospheric profiles. Therefore, perfect alignment between Jacobians from the standard atmosphere and those at the perturbed location is impractical.

Ideally, Jacobians should be derived from local temperature and humidity profiles at the perturbed location, but this is not feasible with the output of FuXi Weather, which provides only 13 pressure levels and lacks the specific humidity, cloud-water, and cloud-ice data required for precise Jacobian derivation using RTTOV. Therefore, the standard US atmospheric profile was employed for illustrative purposes. Supplementary Fig. 11 provides the temperature and humidity Jacobian functions derived from the ATMS aboard NOAA-20.



Supplementary Figure 11: Jacobian functions for temperature (T) and humidity (Q) derived from Advanced Technology Microwave Sounder (ATMS) aboard NOAA-20. The left panel shows T Jacobians for all channels, while the right panel displays Q Jacobians for all channels. The atmospheric profile is based on the US standard atmosphere, and radiative transfer calculations are performed using RTTOV version 12.2.

5 Data denial experiments

In the main text, FuXi Weather was assessed based on experiments that assimilate the full set of satellite data, which will be referred to as the control run. To measure the impact of excluding specific satellite observations on global analysis and forecasting, two sets of data denial experiments were conducted;

this is a common approach for understanding the contributions of different observations in DA systems [3].

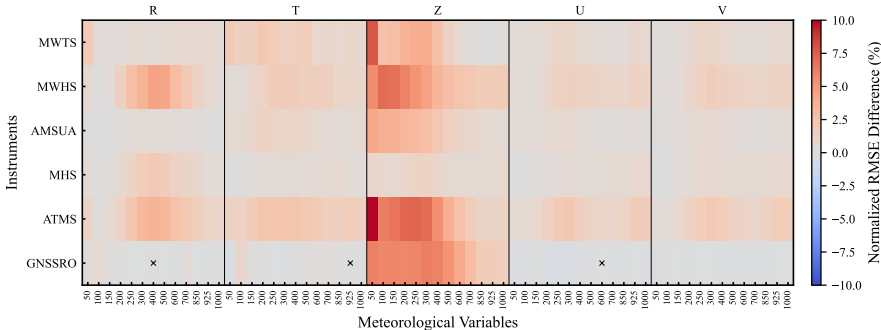
In the first set of experiments, 6-hour forecasts from the control run served as background fields at each analysis time, with specific satellite data excluded during DA. The resulting analysis fields were compared with those from the control run. In the second set of experiments, background forecasts from the control run were not used, and specific satellite data were withheld for DA. The generated analysis fields were used for new 6-hour forecasts, which then became the background fields. This setup provided insights into the robustness of FuXi Weather by examining error evolution. Unlike the first set of experiments, which use identical background fields across all runs, the second set employs cycled background fields that evolve differently depending on the excluded satellite data. Both sets of experiments quantified the impact by calculating the percentage change in RMSE, using the control run as the baseline. In these experiments, the data from each satellite instrument are excluded separately.

Supplementary Fig. 12 illustrates the normalized differences in the globally-averaged and latitude-weighted RMSE of the analysis fields. The differences are predominantly positive, as shown by red shading, suggesting excluding any satellite observation generally leads to reduced accuracy compared to the control run. Denying MWHS and ATMS led to the largest accuracy reductions for R, because these instruments provide global temperature and humidity soundings under all weather conditions. In contrast, excluding MHS had a negligible impact owing to its lower resolution and fewer channels. The temperature was most affected by removing MWTS, MWHS, AMSU-A, and ATMS, which all included temperature-sounding channels. The U and V wind components were sensitive to the denial of humidity sounders, particularly MWHS and ATMS, owing to the "generalized tracer effect". Geopotential suffered the greatest degradation, with RMSE values increasing by 20% at multiple pressure levels, especially when MWHS, ATMS, and GNSS-RO data were excluded, highlighting the sensitivity of geopotential to accurate representations of temperature, humidity, and wind. It is noteworthy that the denial of ATMS primarily affected R below 200 hPa, consistent with its Jacobian functions in Supplementary Fig. 11, where values are close to zero above 200 hPa. The impact of MWTS on T was greater at higher altitudes, consistent with its greater Jacobian values at these altitudes.

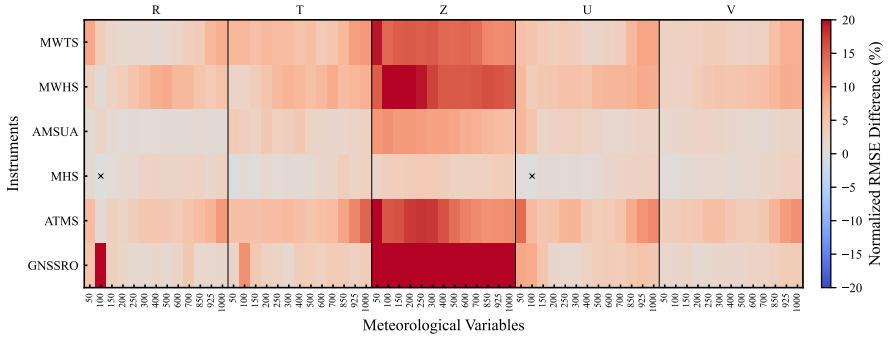
Supplementary Fig. 13 presents results from the second set of experiments, involving more complex variable correlations. Removing specific satellite data degraded not only the analysis fields but also the subsequent forecasts because errors propagated across variables. For example, while GNSS-RO denial primarily affected geopotential at upper-pressure levels in the first set of experiments, it impacted nearly all variables in this second set. Despite this, FuXi Weather demonstrated robustness, with error growth remaining within acceptable limits over the 1-year testing period.

Overall, MWTS and MWHS aboard FY-3E and ATMS aboard NOAA-20 provided the most influential observations. These data denial experiments are

valuable in assessing the performance of DA systems with limited observations and help to evaluate the relative impacts of different observations as new data sources become available. The results generally aligned with the temperature and humidity Jacobian functions, confirming FuXi Weather’s ability to evaluate the impact of different observational systems on analysis and forecast accuracy.



Supplementary Figure 12: Percentage change in the average errors of analysis fields from data denial experiments. Each experiment uses 6-hour control forecast backgrounds at analysis times, excluding one satellite data type during DA. The figure displays the normalized differences in the globally-averaged and latitude-weighted root mean square error (RMSE) of analysis fields produced by the FuXi Weather system over a 1-year testing period (July 3, 2023, to June 30, 2024). The data from each satellite instrument are excluded individually, and the resulting analysis RMSE is compared with that from the control run, which assimilates all satellite data. The comparison includes five upper-air atmospheric variables across 13 pressure levels (50, 100, 150, 200, 250, 300, 400, 500, 600, 700, 850, 925, and 1000 hPa). The five upper-air atmospheric variables are relative humidity (R), temperature (T), geopotential (Z), u component of wind (U), and v component of wind (V). Red and blue shading indicates degradation and improvement in accuracy compared with the control run, respectively, with varying impacts across variables and pressure levels. The X symbols denote cases where the resulting RMSE is not statistically significant, based on the t-test at the 95% confidence level.



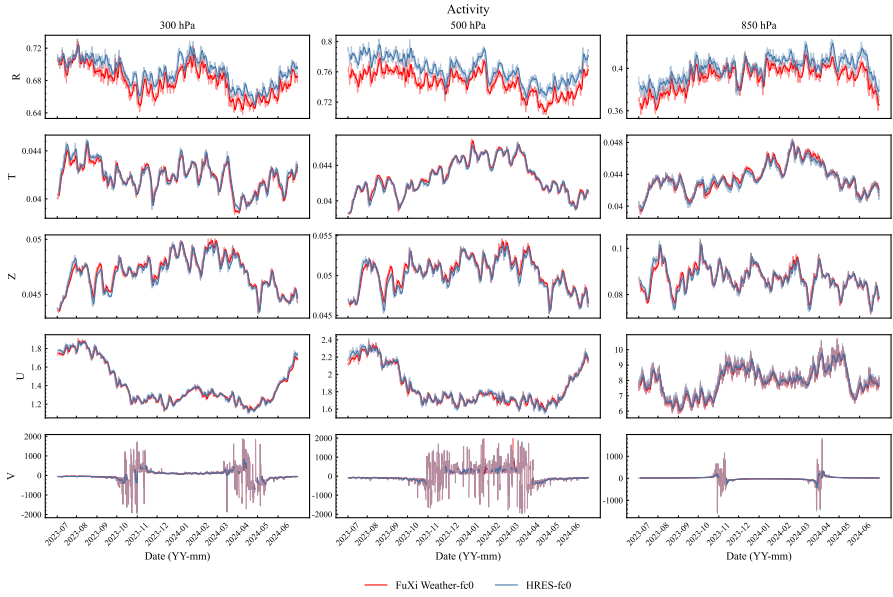
Supplementary Figure 13: Percentage change in the average errors of analysis fields from data denial experiments. Each experiment uses cycled background fields that evolve differently depending on the excluded satellite data. The figure displays the normalized differences in the globally-averaged and latitude-weighted root mean square error (RMSE) of analysis fields produced by the FuXi Weather system over a 1-year testing period (July 3, 2023 - June 30, 2024). The data from each satellite instrument are excluded individually, and the resulting analysis RMSE is compared with that of the control run, which assimilates all satellite data. Red and blue shading indicates degradation and improvement in accuracy compared with the control run, respectively, with varying impacts across variables and pressure levels. The X symbols denote cases where the resulting RMSE is not statistically significant, based on the t-test at the 95% confidence level.

6 Additional evaluations of global analysis fields

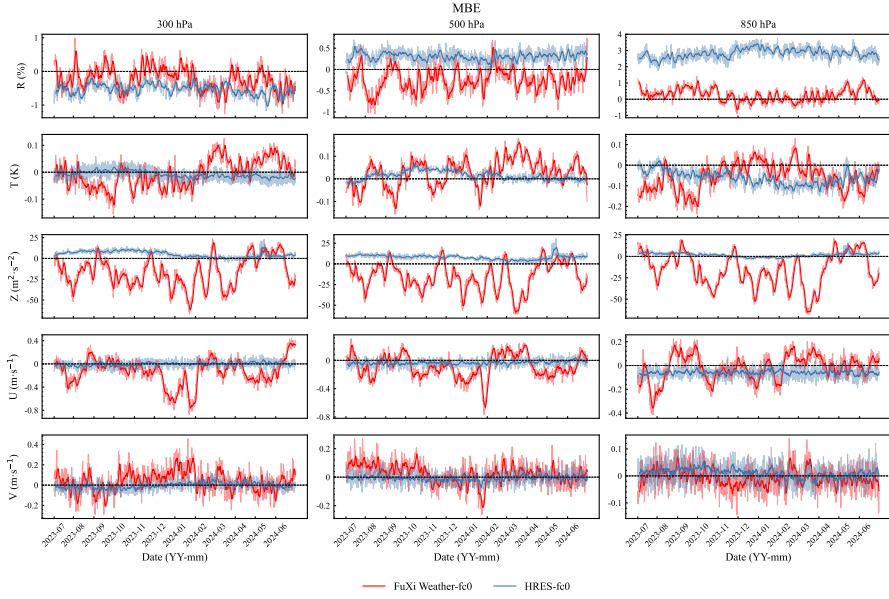
Figure 14 compares the analysis activity (as defined in Equation 7 in Section 4.2 of the main text) between FuXi Weather and ECMWF HRES over a 1-year testing period. While both systems show similar temporal patterns for most variables, FuXi Weather consistently exhibits slightly lower analysis activity values for R. Notably, the V wind component demonstrates significantly higher activity values than other variables due to its near-zero climatological mean, which amplifies relative variability in the metric.

Figure 15 presents the 1-year time series of MBE for both systems relative to ERA5. For T and wind components (U, V), MBE values cluster near zero in both systems, indicating minimal systematic bias in FuXi Weather. However, distinct bias patterns emerge for R: ECMWF HRES shows consistent negative biases at 300 hPa and positive biases at 500 and 850 hPa, while FuXi Weather maintains near-zero MBE values across all levels. For the variable Z, FuXi Weather analyses demonstrate substantially larger errors than ECMWF HRES. Overall, FuXi Weather shows greater temporal variability, likely due to the fewer assimilated observations and lack of quality control compared to ECMWF HRES.

Allen et al. [49] introduced Aardvark Weather, an end-to-end data-driven weather prediction system that does not use background forecasts during DA. Their findings, presented in Figure 2 of the main text and Figure 1 of the Supplementary Information, indicate that Aardvark Weather analyses (at lead time $t=0$) achieve RMSE values comparable to those of ECMWF HRES forecasts at a 2-day lead time. When comparing the performance of Aardvark and FuXi Weather, it is important to note several key differences. Aardvark operates at a coarser spatial resolution of 1.5° and a temporal resolution of 24 hours, while FuXi Weather uses a finer spatial resolution of 0.25° and a 6-hour temporal resolution. Additionally, Aardvark assimilates satellite observations over a 24-hour DA window, in contrast to FuXi Weather’s current 8-hour DA window. Furthermore, Aardvark’s performance is compared with ECMWF HRES forecasts, which exhibit slightly higher RMSE values than FuXi forecasts (initialized from ERA5) at the same lead times.



Supplementary Figure 14: Comparison of analysis activity produced by FuXi Weather and ECMWF HRES over a 1-year testing period from July 03, 2023, to June 30, 2024. The time series shows the globally-averaged and latitude-weighted analysis activity for: the analysis fields of FuXi Weather with background forecasts (solid red lines) and ECMWF HRES (solid blue lines). The comparison includes five variables: relative humidity (R), temperature (T), geopotential (Z), u component of wind (U), and v component of wind (V), at three pressure levels (300 hPa, 500 hPa, and 850 hPa). The five rows and three columns correspond to five variables and three pressure levels, respectively. To improve clarity, the original data are shown with reduced opacity, while solid lines represent smoothed values using a 12-point moving average.



Supplementary Figure 15: Comparison of analysis mean bias error (MBE) produced by FuXi Weather and ECMWF HRES over a 1-year testing period from July 03, 2023, to June 30, 2024. The time series shows the globally-averaged and latitude-weighted MBE relative to ERA5 for: the analysis fields of FuXi Weather with background forecasts (solid red lines) and ECMWF HRES (solid blue lines). The comparison includes five variables: relative humidity (R), temperature (T), geopotential (Z), u component of wind (U), and v component of wind (V), at three pressure levels (300 hPa, 500 hPa, and 850 hPa). The five rows and three columns correspond to five variables and three pressure levels, respectively. To improve clarity, the original data are shown with reduced opacity, while solid lines represent smoothed values using a 12-point moving average.

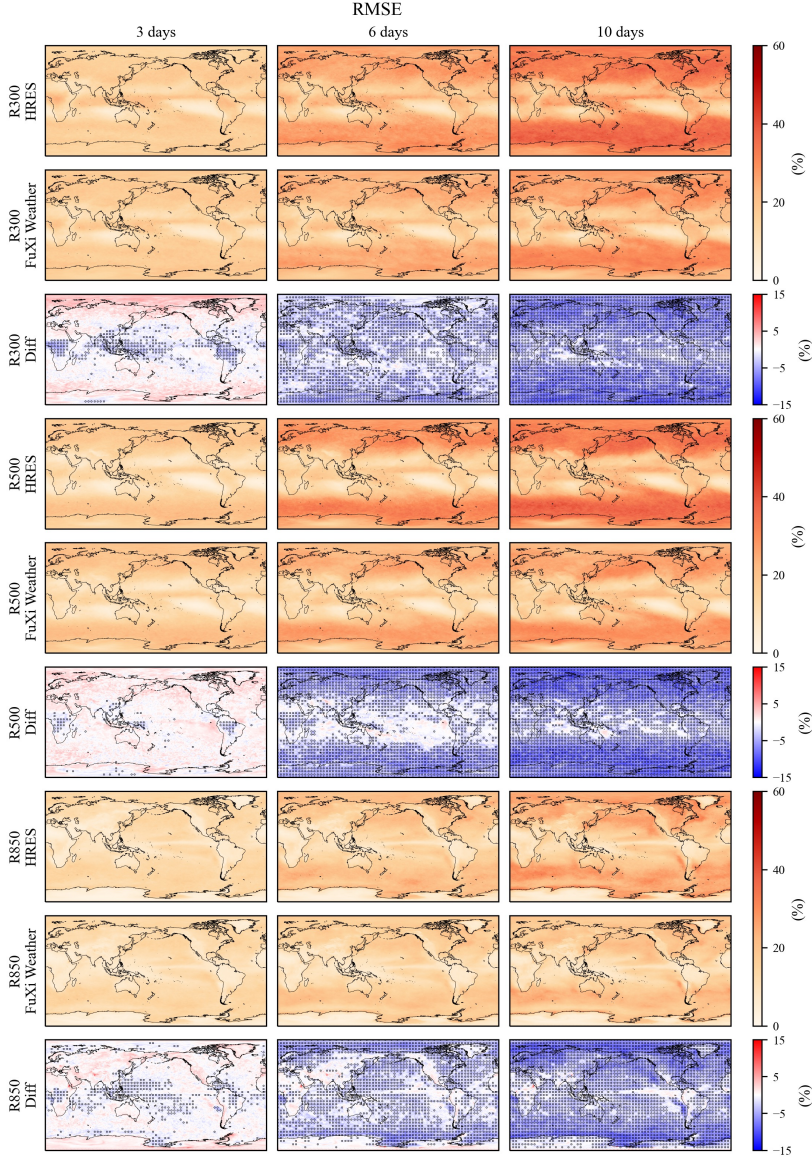
7 Spatial distributions of global weather forecast performance

7.1 Forecasts for upper-air variables

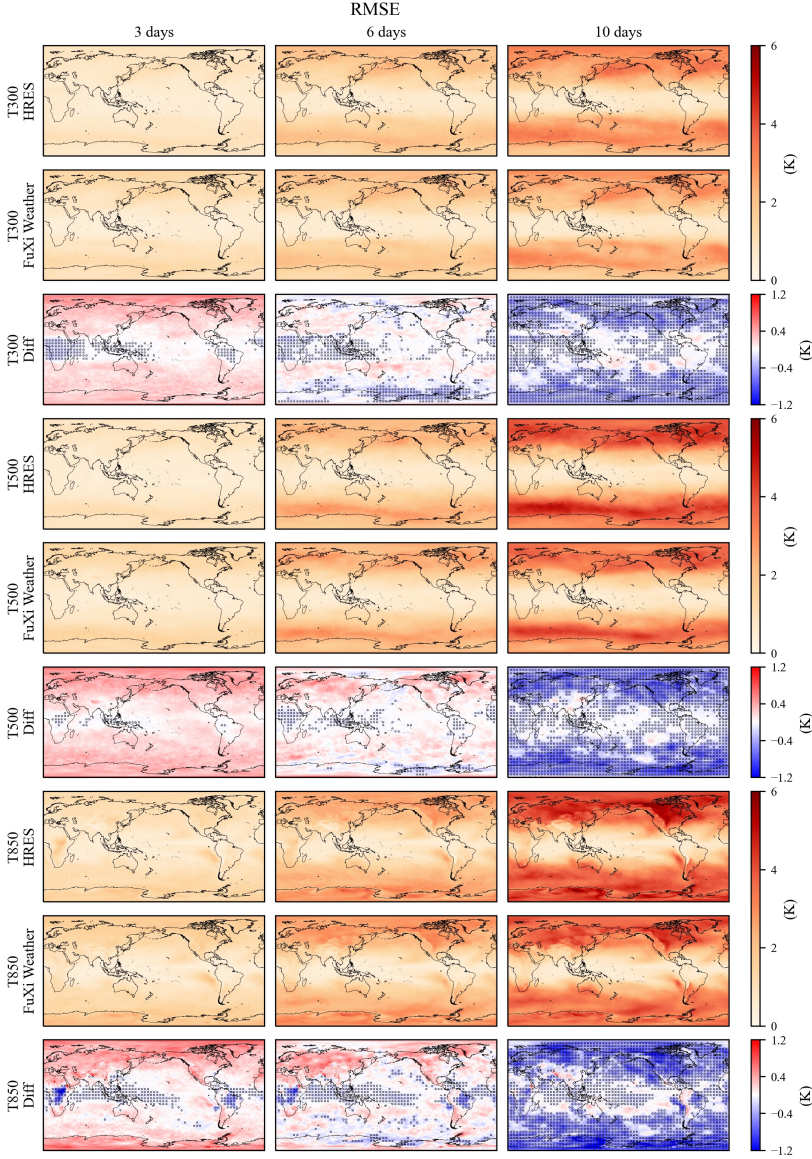
Supplementary Figs. 16-20 present spatial comparisons of the average RMSE without latitude weighting for forecasts from ECMWF HRES and FuXi. These figures illustrate RMSE and values for three pressure levels (300, 500, and 850 hPa) and five upper-air variables: R, T, Z, U, and V. Forecast lead times of 3, 6, and 10 days are represented in the first, second, and third columns of each figure, respectively. Darker red shades indicate higher RMSE values, showing a general increase in error with longer lead times, particularly in extra-tropical regions.

Spatial maps of RMSE differences reveal areas where FuXi outperforms ECMWF (blue), where ECMWF is superior (red), and where the two performances are comparable (white). At a 3-day lead time, red dominates the maps for the variables T, Z, U, and V, suggesting that ECMWF is superior, especially in extra-tropical regions. However, for R, FuXi shows comparable or better performance at 300 and 850 hPa, aligning with the latitude-weighted RMSE analysis in the main text Fig. 2. At a 10-day lead time, most regions exhibit predominantly blue or white colors, suggesting an overall better performance by FuXi, particularly in extratropical areas. The intermediate 6-day forecast shows a more prominent shift to blue for variable R than for the four other variables, indicating the enhanced performance of FuXi for this variable among all evaluated upper-air parameters.

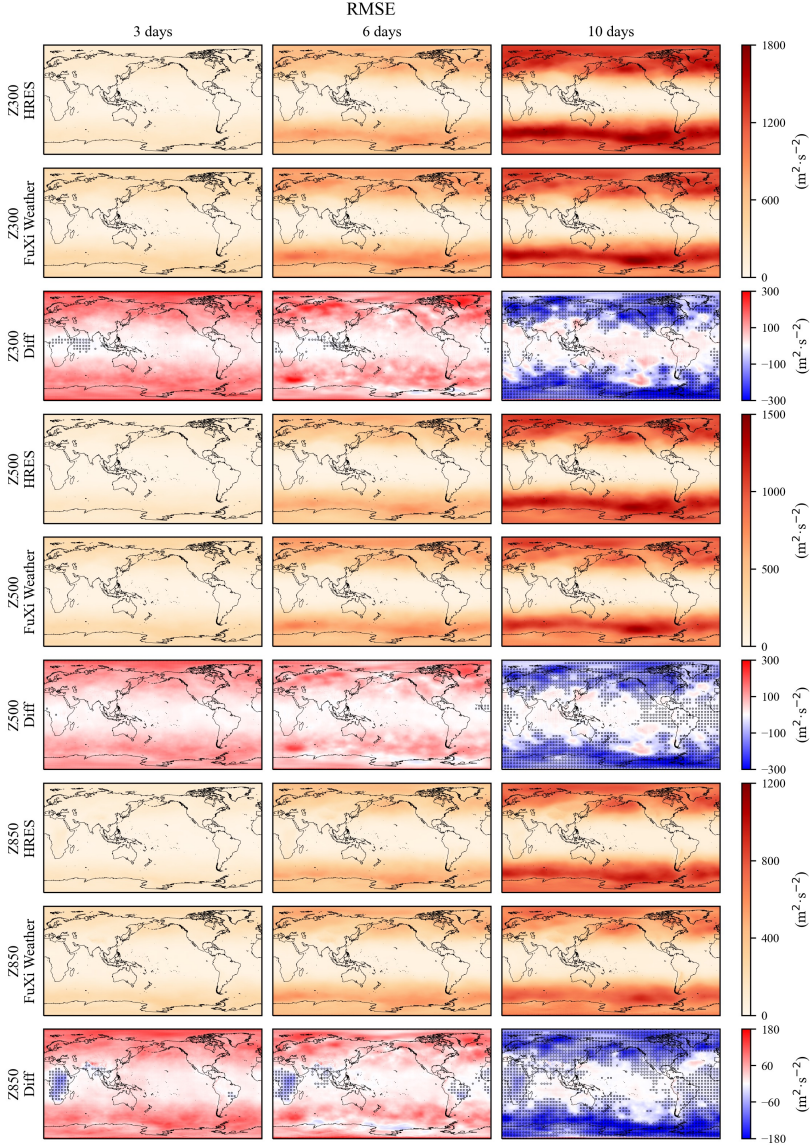
Notably, FuXi Weather consistently demonstrates lower RMSE values for R, T, and Z across all lead times over central Africa, suggesting better performance in this region compared to other parts of the world. This is likely due to the limited availability of conventional observations in Africa, which restricts the data available for ECMWF HRES assimilation. Consequently, satellite data plays a more prominent role in Africa. Furthermore, because of the predominantly westerly winds and high topography of eastern Africa, including prominent mountain ranges, these improvements tend to propagate westwards toward South America rather than eastwards. As a result, forecast accuracy in South America tends to increase with longer forecast lead times. The superior performance of FuXi Weather also propagates to the North and South Poles, with this effect being especially noticeable at a lead time of 10 days.



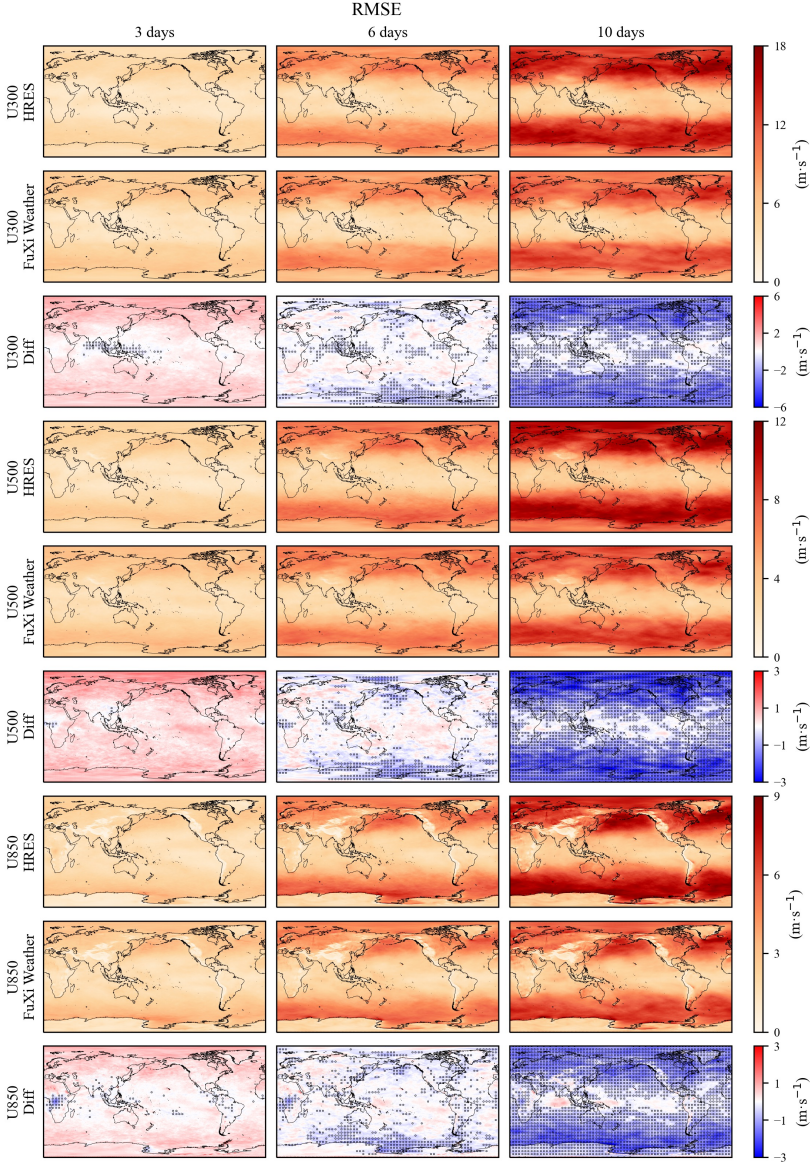
Supplementary Figure 16: Spatial maps of average root mean square error (RMSE) without latitude weighting of forecasts from ECMWF HRES (first, fourth, and seventh rows) and FuXi (second, fifth, and eighth rows), along with the differences in RMSE (third, sixth, and ninth rows) between FuXi and ECMWF HRES for relative humidity (R) across three pressure levels (300, 500, and 850 hPa). The maps correspond to forecast lead times of 3 days (first column), 6 days (second column), and 10 days (third column), using all testing data over a 1-year testing period, spanning July 03, 2023-June 30, 2024. In the RMSE maps, darker red indicates higher RMSE values. In the difference maps, blue and red colors denote regions where FuXi or ECMWF HRES performs better, respectively. Stippling on the maps in the third, sixth, and ninth rows highlights regions where FuXi Weather significantly outperforms ECMWF HRES, based on the t-test at the 95% confidence level.



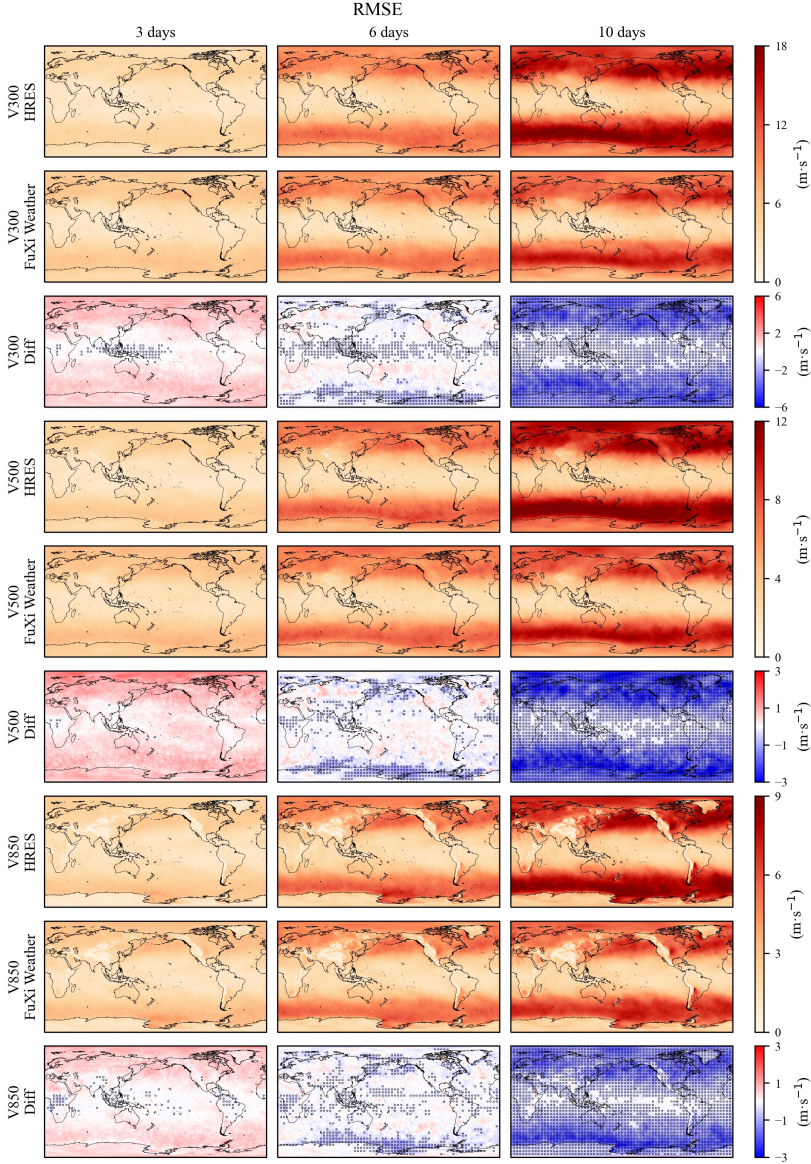
Supplementary Figure 17: Spatial maps of average root mean square error (RMSE) without latitude weighting of forecasts from ECMWF HRES (first, fourth, and seventh rows) and FuXi (second, fifth, and eighth rows), along with differences in RMSE (third, sixth, and ninth rows) between FuXi and ECMWF HRES for temperature (T) across three pressure levels (300, 500, and 850 hPa). The maps correspond to forecast lead times of 3 days (first column), 6 days (second column), and 10 days (third column), using all testing data over a 1-year testing period, spanning July 03, 2023-June 30, 2024. In the RMSE maps, darker red indicates higher RMSE values. In the difference maps, blue and red colors denote regions where FuXi or ECMWF HRES performs better, respectively. Stippling on the maps in the third, sixth, and ninth rows highlights regions where FuXi Weather significantly outperforms ECMWF HRES, based on the t-test at the 95% confidence level.



Supplementary Figure 18: Spatial maps of average root mean square error (RMSE) without latitude weighting of forecasts from ECMWF HRES (first, fourth, and seventh rows) and FuXi (second, fifth, and eighth rows), along with the differences in RMSE (third, sixth, and ninth rows) between FuXi and ECMWF HRES for geopotential (Z) across three pressure levels (300, 500, and 850 hPa). The maps correspond to forecast lead times of 3 days (first column), 6 days (second column), and 10 days (third column), using all testing data over a 1-year testing period, spanning July 03, 2023-June 30, 2024. In the RMSE maps, darker red indicates higher RMSE values. In the difference maps, blue and red colors denote regions where FuXi or ECMWF HRES performs better, respectively. Stippling on the maps in the third, sixth, and ninth rows highlights regions where FuXi Weather significantly outperforms ECMWF HRES, based on the t-test at the 95% confidence level.



Supplementary Figure 19: Spatial maps of average root mean square error (RMSE) without latitude weighting of forecasts from ECMWF HRES (first, fourth, and seventh rows) and FuXi (second, fifth, and eighth rows), along with the differences in RMSE (third, sixth, and ninth rows) between FuXi and ECMWF HRES for the u component of wind (U) across three pressure levels (300, 500, and 850 hPa). The maps correspond to forecast lead times of 3 days (first column), 6 days (second column), and 10 days (third column), using all testing data over a 1-year testing period, spanning July 03, 2023–June 30, 2024. In the RMSE maps, darker red indicates higher RMSE values. In the difference maps, blue and red colors denote regions where FuXi or ECMWF HRES performs better, respectively. Stippling on the maps in the third, sixth, and ninth rows highlights regions where FuXi Weather significantly outperforms ECMWF HRES, based on the t-test at the 95% confidence level.

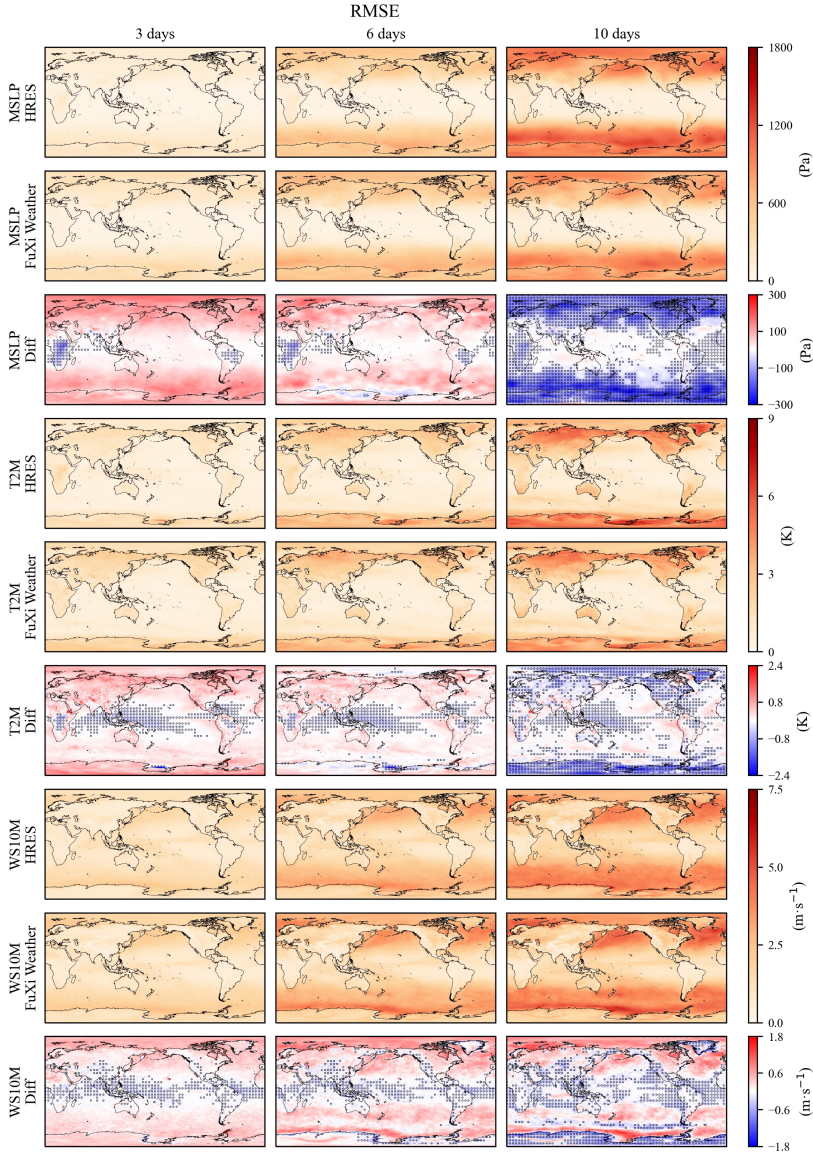


Supplementary Figure 20: Spatial maps of average root mean square error (RMSE) without latitude weighting of forecasts from ECMWF HRES (first, fourth, and seventh rows) and FuXi (second, fifth, and eighth rows), along with the differences in RMSE (third, sixth, and ninth rows) between FuXi and ECMWF HRES for the v component of wind (V) across three pressure levels (300, 500, and 850 hPa). The maps correspond to forecast lead times of 3 days (first column), 6 days (second column), and 10 days (third column), using all testing data over a 1-year testing period, spanning July 03, 2023–June 30, 2024. In the RMSE maps, darker red indicates higher RMSE values. In the difference maps, blue and red colors denote regions where FuXi or ECMWF HRES performs better, respectively. Stippling on the maps in the third, sixth, and ninth rows highlights regions where FuXi Weather significantly outperforms ECMWF HRES, based on the t -test at the 95% confidence level.

7.2 Forecasts for surface variables

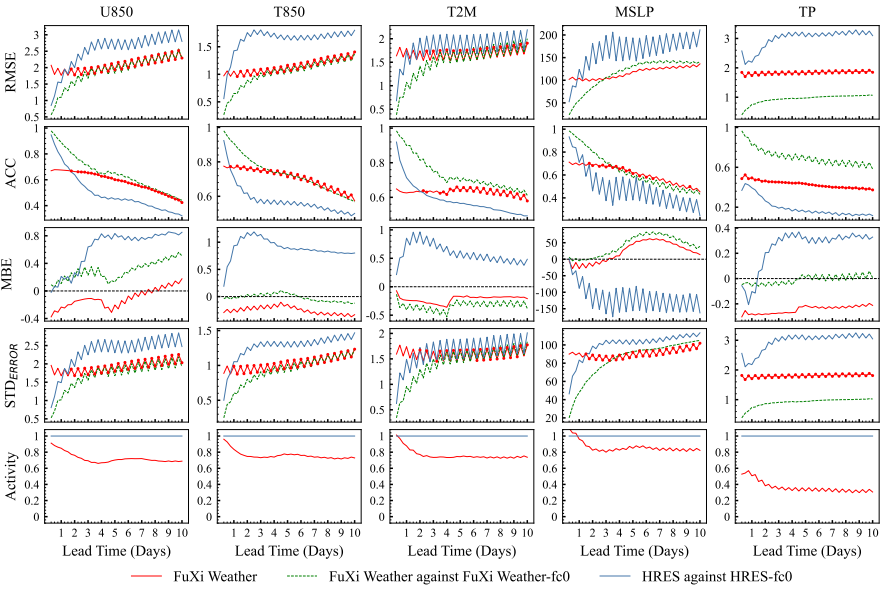
Supplementary Fig. 21 presents spatial comparisons of the average RMSE, without latitude weighting, for ECMWF HRES and FuXi forecasts of surface variables, including MSLP, T2M, and 10-meter wind speed (WS10M). Consistent with the results for upper-air variables, the figure shows that RMSE differences are predominantly negative over central Africa at forecast lead times of 3, 6, and 10 days, indicating FuXi Weather’s superior performance.

Additionally, FuXi Weather outperforms ECMWF HRES in other regions where land-based observations are sparse, such as India and areas near the North and South Poles. This superior performance becomes more pronounced as the forecast lead time extends from 6 to 10 days, with an increasing area showing negative RMSE differences, suggesting FuXi Weather’s growing advantage in forecast accuracy over longer lead times.

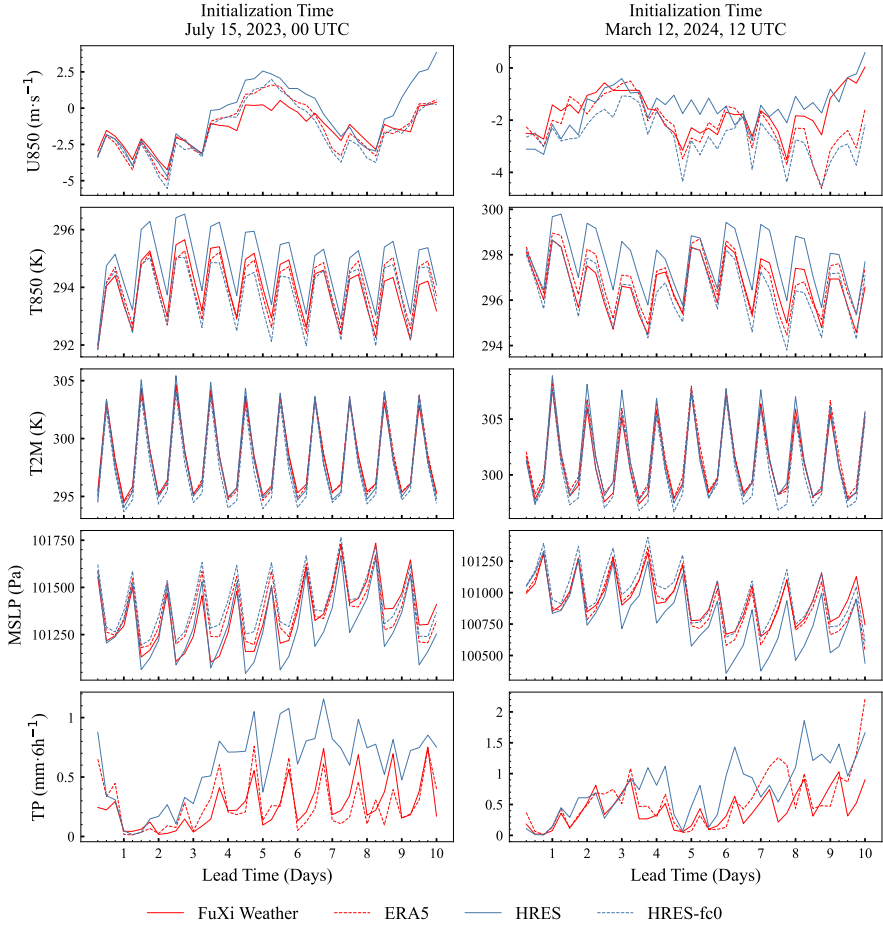


Supplementary Figure 21: Spatial maps of average root mean square error (RMSE) without latitude weighting of forecasts from ECMWF HRES (first, fourth, and seventh rows) and FuXi (second, fifth, and eighth rows), along with the differences in RMSE (third, sixth, and ninth rows) between FuXi and ECMWF HRES for mean sea-level pressure (MSLP), 2-meter temperature (T2M), and 10-meter wind speed (WS10M). The maps correspond to forecast lead times of 3 days (first column), 6 days (second column), and 10 days (third column), using all testing data over a 1-year testing period, spanning July 03, 2023-June 30, 2024. In the RMSE maps, darker red indicates higher RMSE values. In the difference maps, blue and red colors denote regions where FuXi or ECMWF HRES performs better, respectively. Stippling on the maps in the third, sixth, and ninth rows highlights regions where FuXi Weather significantly outperforms ECMWF HRES, based on the t-test at the 95% confidence level.

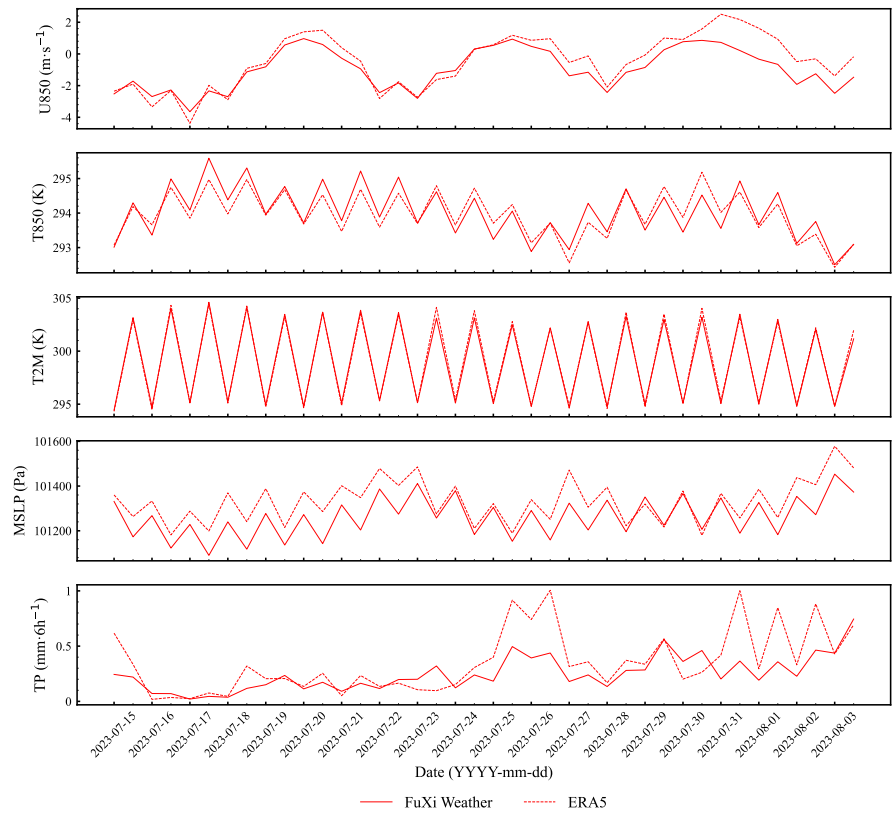
8 Forecast performance in central Africa and South America



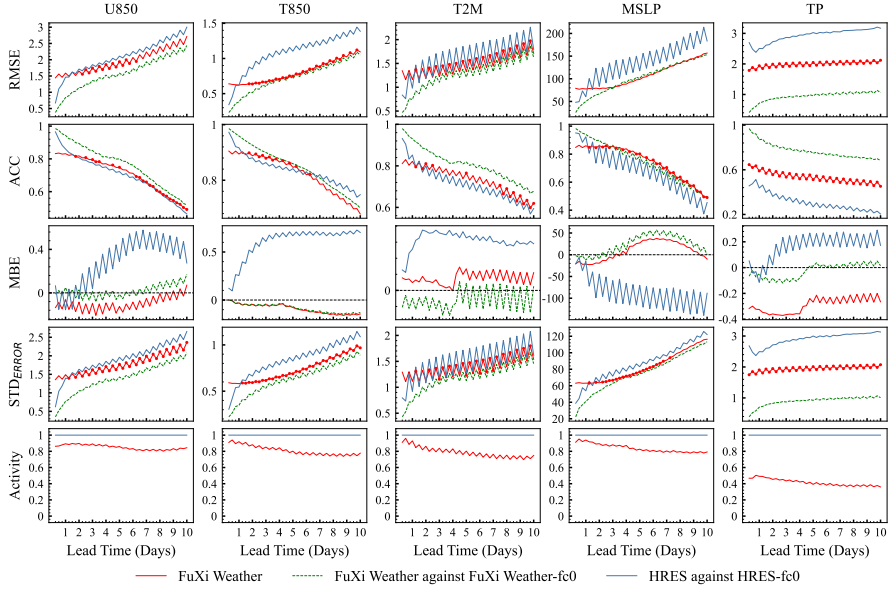
Supplementary Figure 22: Comparison of forecast performance over central Africa during the 1-year testing period from July 03, 2023, to June 30, 2024. Central Africa is defined as the region spanning 15° E to 35° E in longitude and 10° N to 10° S in latitude. Rows 1 to 5 show the root mean square error (RMSE), anomaly correlation coefficient (ACC), mean bias error (MBE), standard deviation of errors (STD_{ERROR}), and activity for forecasts generated by FuXi Weather (red solid and green dashed lines) and ECMWF HRES (blue). FuXi Weather is initialized using analysis fields produced by FuXi-DA incorporating background forecasts. This figure includes five variables: 850 hPa u wind component (U850), 850 hPa temperature (T850), 2-meter temperature (T2M), mean sea level pressure (MSLP), and total precipitation (TP; units mm). FuXi forecasts (red lines) are verified against ERA5, and also against FuXi-DA analyses (green dashed lines). When FuXi (green dashed lines) and ECMWF HRES (blue) forecasts are evaluated against their respective initialization time series, they inherently exhibit lower error and higher ACC in early lead times. Red dots indicate time steps where FuXi Weather significantly outperforms ECMWF HRES, based on the t-test at the 95% confidence level.



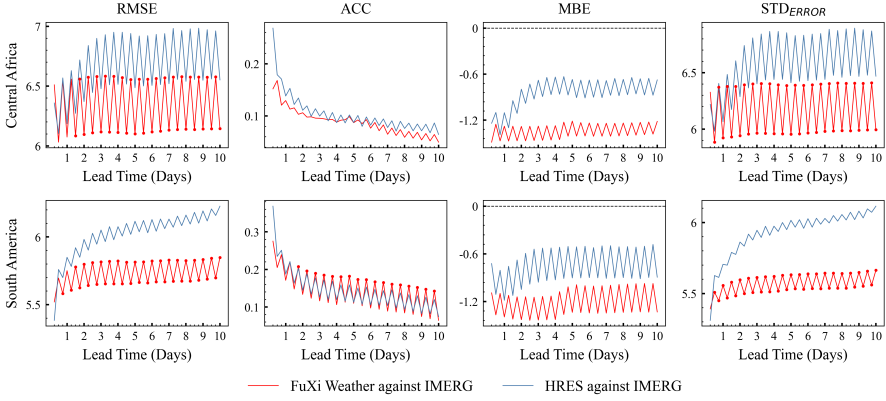
Supplementary Figure 23: Example time series of 10-day forecasts averaged over central Africa for two initialization times: July 15, 2023, 00 UTC (left column), and March 12, 2024, 12 UTC (right column). Central Africa is defined as the region spanning 15° E to 35° E in longitude and 10° N to 10° S in latitude. The forecasts include FuXi Weather (solid red lines) and its benchmark ERA5 (dashed red lines), as well as ECMWF HRES (solid blue lines) and its benchmark HRES-fc0 (dashed blue lines). Rows 1 to 5 show the time series for 850 hPa u wind component (U850, first row), 850 hPa temperature (T850, second row), 2-meter temperature (T2M, third row), mean sea level pressure (MSLP, fourth row), and total precipitation (TP, fifth row; units mm).



Supplementary Figure 24: Example time series of 3-day lead time forecasts averaged over central Africa from July 15, 2023, to August 3, 2023. Central Africa is defined as the region spanning 15° E to 35° E in longitude and 10° N to 10° S in latitude. The forecasts include FuXi Weather (solid red lines) and its benchmark ERA5 (dashed red lines). Rows 1 to 5 display the time series for 850 hPa u wind component (U850, first row), 850 hPa temperature (T850, second row), 2-meter temperature (T2M, third row), mean sea level pressure (MSLP, fourth row), and total precipitation (TP, fifth row; units mm).



Supplementary Figure 25: Comparison of forecast performance over northern South America during the 1-year testing period from July 03, 2023, to June 30, 2024. Northern South America is defined as the region spanning 70° W to 50° W in longitude and 0° S to 20° S in latitude. Rows 1 to 5 show the root mean square error (RMSE), anomaly correlation coefficient (ACC), mean bias error (MBE), and standard deviation of errors (STD_{ERROR}), and activity for forecasts generated by FuXi Weather (red solid and green dashed lines) and ECMWF HRES (blue). FuXi Weather is initialized using analysis fields produced by FuXi-DA incorporating background forecasts. This figure includes five variables: 850 hPa u wind component (U850), 850 hPa temperature (T850), 2-meter temperature (T2M), mean sea level pressure (MSLP), and total precipitation (TP; units mm). FuXi forecasts (red lines) are verified against ERA5, and also against FuXi-DA analyses (green dashed lines). When FuXi (green dashed lines) and ECMWF HRES (blue) forecasts are evaluated against their respective initialization time series, they inherently exhibit lower error and higher ACC in early lead times. Red dots indicate time steps where FuXi Weather significantly outperforms ECMWF HRES, based on the t-test at the 95% confidence level.



Supplementary Figure 26: Comparison of total precipitation (TP; units mm) forecast performance over central Africa (top row) and northern South America (bottom row) during the 1-year testing period from July 03, 2023, to June 30, 2024. Central Africa is defined as the region spanning 15° E to 35° E in longitude and 10° N to 10° S in latitude. Northern South America is defined as the region spanning 70° W to 50° W in longitude and 0° S to 20° S in latitude. Columns 1 to 4 show the root mean square error (RMSE), anomaly correlation coefficient (ACC), mean bias error (MBE), and standard deviation of errors (STD_{ERROR}) for forecasts generated by FuXi Weather (red) and ECMWF HRES (blue). FuXi Weather is initialized using analysis fields produced by FuXi-DA incorporating background forecasts. Both FuXi and ECMWF HRES forecasts are verified against IMERG instead of ERA5. Red dots indicate time steps where FuXi Weather significantly outperforms ECMWF HRES, based on the t-test at the 95% confidence level.

9 Importance of incorporating background forecasts in the data-to-forecast system

In DA, challenges often arise due to incomplete, noisy, or sparse observational data, as well as uncertainties in the forecasting models. Background forecasts, also referred to as "first guess" fields, play a critical role in DA by providing prior estimates of the atmospheric state. They serve as a form of regularization, constraining the solution space and mitigating overfitting to noisy or sparse observations. In variational DA methods such as 3D-Var and 4D-Var, the cost function includes a term that penalizes deviations from the background forecast, effectively balancing observational data with model predictions [50].

In the main text, we compare two variants of FuXi-DA, one with and one without background forecasts, to assess the importance of background forecasts in machine learning-based DA. The two variants of FuXi-DA are trained independently but share the same model architecture described in Section 2.2. The variant of FuXi Weather without background forecasts relies solely on observational data. Consequently, it does not involve cycling, as forecasts are generated directly from analysis fields produced by FuXi-DA without incorporating prior forecast information. Figure 1 of the main text demonstrates that analysis fields generated without background forecasts exhibit significantly higher RMSE values throughout the entire one-year testing period. Figures 2 and 3 of the main text further compare the RMSE and ACC of forecasts derived from analysis fields with and without background forecasts, reinforcing the conclusion that background forecasts are essential for improving the performance of machine learning-based DA.

10 Effect of different training settings

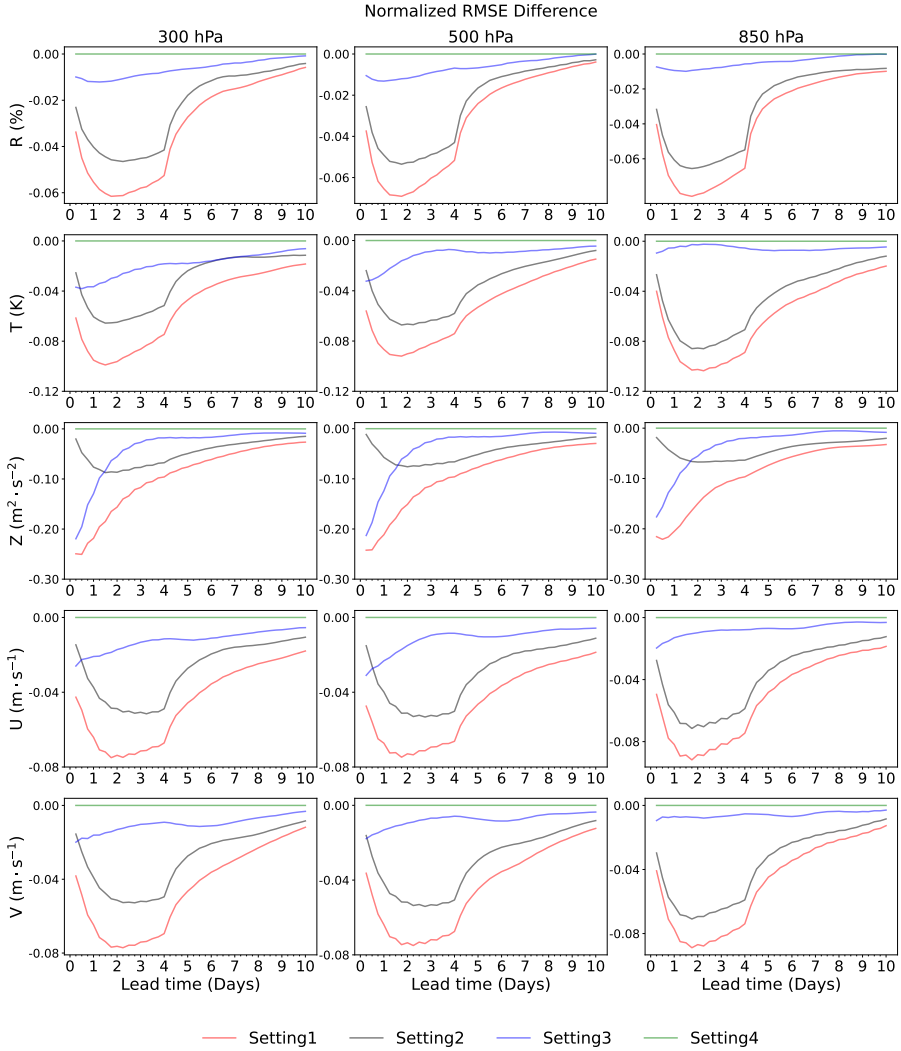
This section evaluates the performance of FuXi Weather under four training configurations, detailed in Supplementary Table 2. Corresponding to the FuXi Weather configuration discussed in the main text, setting 1 includes incremental learning in FuXi-DA and fine-tuning of FuXi-Short. Setting 2 excludes incremental learning, setting 3 omits the fine-tuning of FuXi-Short, and setting 4 excludes both.

Supplementary Figs. 27 and 28 present the normalized differences in the globally-averaged and latitude-weighted RMSE and ACC for the four models, using setting 4 as the baseline. The results indicate that setting 1, which incorporates incremental learning and fine-tuning, yields the best performance. A comparison between settings 2 and 3 reveals that fine-tuning yields greater improvements than incremental learning alone.

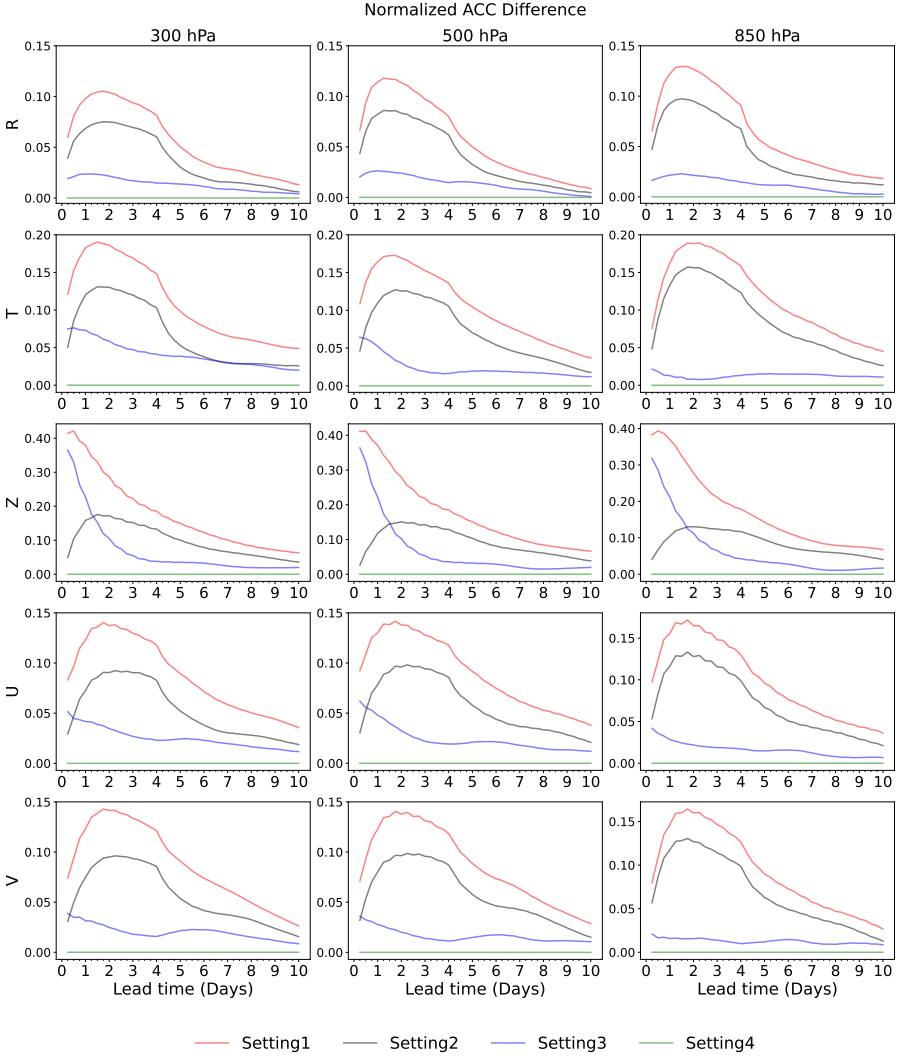
Supplementary Fig. 29 further compares skillful forecast lead times across the four models, supporting the findings described above. The model trained with setting 1 achieves the longest skillful lead times across all 15 variable and pressure-level combinations, confirming the benefits of combining incremental learning with fine-tuning.

Supplementary Table 2: Overview of training settings for FuXi Weather. The settings vary depending on whether incremental learning is implemented in FuXi-DA and whether the FuXi-Short forecasting model is fine-tuned. Setting 1, which includes incremental learning in FuXi-DA and fine-tuning of FuXi-Short, corresponds to the FuXi Weather configuration discussed in the main text.

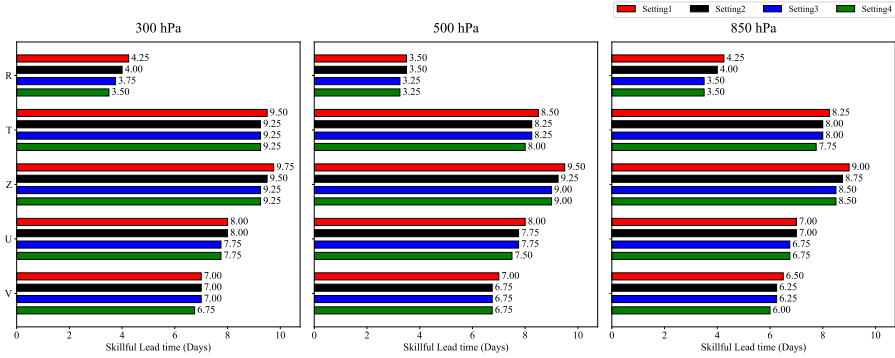
Setting	Incremental learning	Fine-tuning of FuXi-Short
Setting 1	✓	✓
Setting 2	✗	✓
Setting 3	✓	✗
Setting 4	✗	✗



Supplementary Figure 27: Comparison of normalized differences in the globally-averaged and latitude-weighted root mean square error (RMSE) across four FuXi Weather training settings. The settings differ based on the implementation of incremental learning in FuXi-DA and whether the FuXi-Short forecasting model is fine-tuned (see Supplementary Table 2). The comparison spans a 1-year testing period, spanning July 03, 2023-June 30, 2024. The analysis includes five variables: relative humidity (R), temperature (T), geopotential (Z), u component of wind (U), and v component of wind (V), at three pressure levels (300, 500, and 850 hPa). The five rows and three columns correspond to the five variables and three pressure levels, respectively. The performance change on day 4 arises from the model transition from FuXi-Short to FuXi-Medium.



Supplementary Figure 28: Comparison of normalized differences in the globally-averaged and latitude-weighted anomaly correlation coefficient (ACC) across four FuXi Weather training settings. The settings differ based on the implementation of incremental learning in FuXi-DA and whether the FuXi-Short forecasting model is fine-tuned (see Supplementary Table 2). The comparison spans a 1-year testing period, spanning July 03, 2023-June 30, 2024. The analysis includes five variables: relative humidity (R), temperature (T), geopotential (Z), u component of wind (U), and v component of wind (V), at three pressure levels (300 hPa, 500 hPa, and 850 hPa). The five rows and three columns correspond to the five variables and three pressure levels, respectively. The performance change on day 4 arises from the model transition from FuXi-Short to FuXi-Medium.



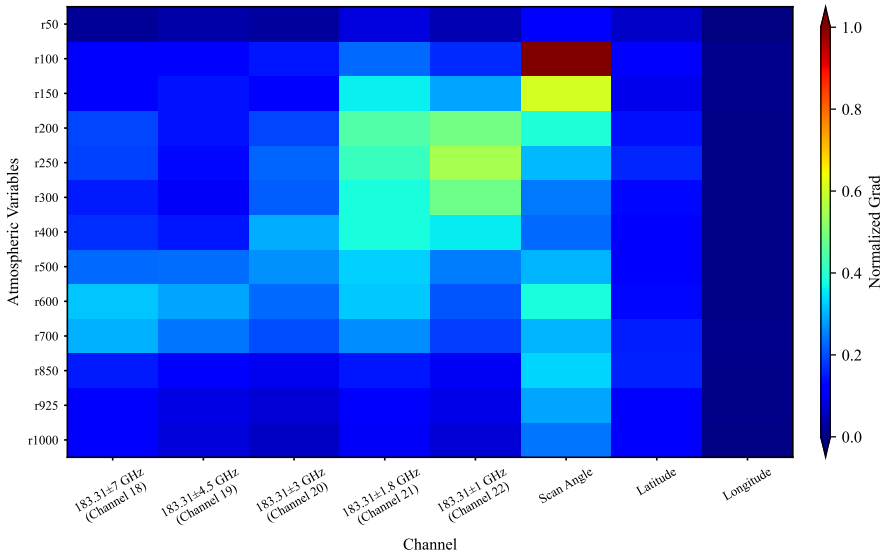
Supplementary Figure 29: Comparisons of skillful forecast lead times across four FuXi Weather training settings. The settings differ based on the implementation of incremental learning in FuXi-DA and whether the FuXi-Short forecasting model is fine-tuned (see Supplementary Table 2). Skillful forecast lead time is defined as the anomaly correlation coefficient (ACC) value above the threshold of 0.6. The comparison spans a 1-year testing period, spanning July 03, 2023-June 30, 2024. The analysis includes five variables: relative humidity (R), temperature (T), geopotential (Z), u component of wind (U), and v component of wind (V), at three pressure levels (300, 500, and 850 hPa). The five rows and three columns correspond to the five variables and three pressure levels, respectively.

11 Sensitivity analysis of observation utilization

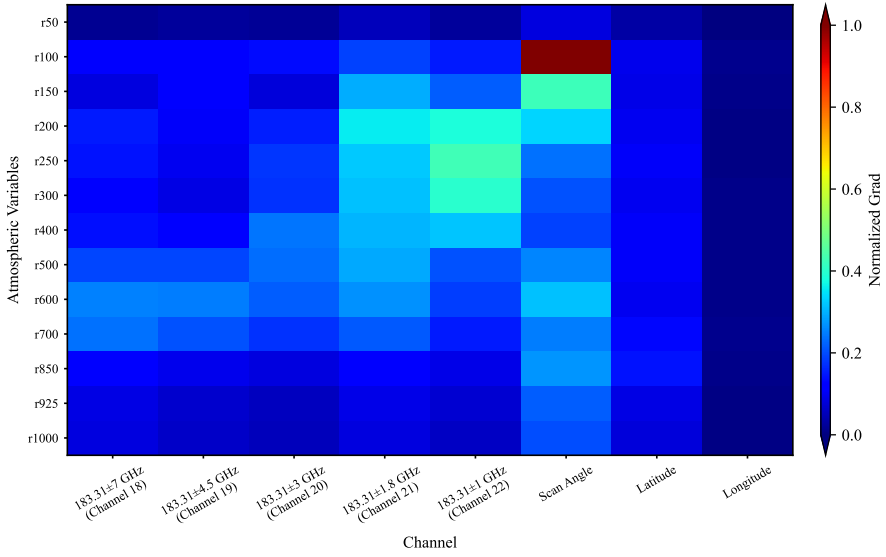
To assess the capability of FuXi-DA in capturing the physical relationships between observations and analysis fields, we conducted a gradient-based sensitivity analysis focusing on the variable R. Specifically, we computed the gradients of R with respect to five humidity-sensitive channels of the NOAA-20 ATMS in the 183.31 GHz band (channels 18-22), as well as scan angle and spatial coordinates (latitude and longitude), for two randomly selected analysis times. At each pressure level, we calculated the latitude-weighted mean absolute error (MAE) loss relative to ERA5 and implemented backward propagation of this MAE loss to obtain the corresponding gradients. The resulting gradients preserved the original dimensionality of ATMS observations ($8 \times 8 \times 720 \times 1440$, where the second 8 represents five channels plus three additional dimensions for scan angle and spatial coordinates). To quantify the magnitude of sensitivity, we took the absolute values of the gradients and aggregated them across the DA temporal window and spatial dimensions (8, 720, 1440), resulting in total gradient magnitudes for each ATMS channel, scan angle, latitude, and longitude. These values were then normalized to the [0,1] range using min-max scaling, with extrema determined across all variables and channels.

Supplementary Figs. 30 and 31 present the resulting sensitivity matrices (13×8) for DA cycles at 06 UTC on July 24, 2023, and 18 UTC on October 2, 2023, respectively. These matrices reveal several key patterns in observation weighting. First, satellite measurements are strongly affected by scan angle. Larger angles correspond to a broader area on the ground, leading to spatial distortion and observational biases [51]. FuXi-DA shows significant sensitivity to scan angle, indicating that the model has adapted to this effect. Second, sensitivity to the scan angle is more pronounced at higher altitudes, particularly near 100 hPa, as reflected by the increased gradient values. This trend is consistent with known characteristics of NOAA-20 ATMS observations, where scan angle effects are greater in the stratosphere (20–50 km). Third, the model shows notable sensitivity to latitude, likely reflecting its ability to account for biases related to airmass, which typically require bias correction in traditional DA methods [52]. In contrast, the relatively minor sensitivity to longitude is expected, as satellite observation accuracy generally exhibits minimal longitudinal variation. Additionally, the five humidity-sensitive channels exhibit distinct vertical sensitivity profiles that vary systematically with frequency offsets (± 1.0 to ± 7.0 GHz). As expected from water vapor absorption physics, channels closer to the line center (e.g., Channel 21 at ± 1.8 GHz and Channel 22 at ± 1.0 GHz) show peak sensitivity at higher altitudes due to stronger absorption. Channels with larger offsets (e.g., Channel 18 at ± 7.0 GHz) penetrate deeper into the atmosphere to sense lower atmospheric layers [53]. This physically consistent pattern is evident in the progressive downward shift in maximum gradient sensitivity with increasing frequency offset.

These results demonstrate that FuXi-DA is capable of learning physically meaningful observation weightings without requiring explicit bias correction. Nevertheless, we acknowledge that integrating established domain knowledge from traditional DA frameworks may further enhance model performance, and we intend to explore this direction in future work.



Supplementary Figure 30: Gradient-based sensitivity analysis of FuXi-DA's relative humidity (R) fields to ATMS observations and encoded observational information during data assimilation at 06 UTC on July 24, 2023. Normalized sensitivity magnitudes (range: 0-1) illustrate the response of R at various pressure levels to: (1) five ATMS humidity-sensitive channels in the 183.31 GHz band (channels 18-22), (2) scan angle, and (3) spatial coordinates (latitude/-longitude). Higher values indicate greater sensitivity.



Supplementary Figure 31: Gradient-based sensitivity analysis of FuXi-DA's relative humidity (R) fields to ATMS observations and encoded observational information during data assimilation at 18 UTC on October 2, 2023. Normalized sensitivity magnitudes (range: 0-1) illustrate the response of R at various pressure levels to: (1) five ATMS humidity-sensitive channels in the 183.31 GHz band (channels 18-22), (2) scan angle, and (3) spatial coordinates (latitude/longitude). Higher values indicate greater sensitivity.

12 Trade-offs between forecast accuracy and activity

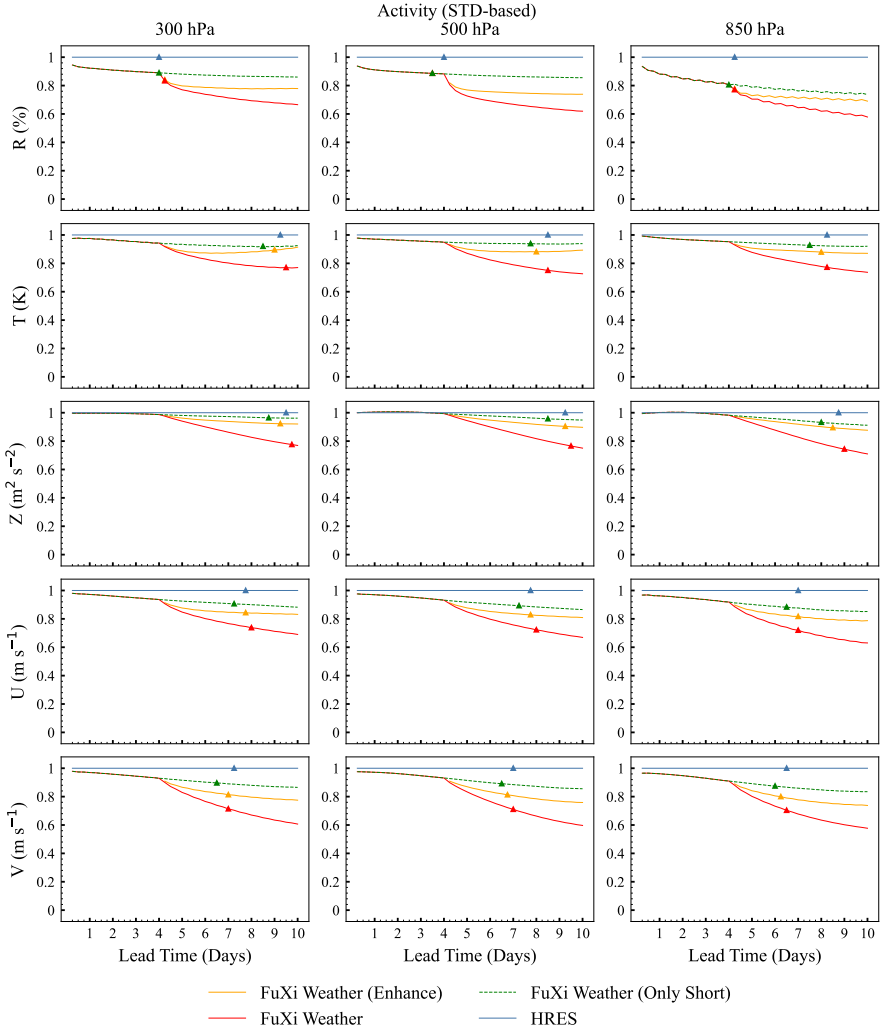
Forecast smoothness in machine learning-based weather predictions primarily arises from two factors. First, these models are typically trained using a mean squared error loss function, which inherently promotes the smoothing of fine-scale features due to the "double penalty" effect [54]. This effect encourages the model to predict the mean of the underlying predictive distribution, thereby suppressing forecast variability. Second, while multi-step fine-tuning for longer forecast lead times can reduce RMSE at longer lead times, it often exacerbates the smoothing effect even more during long autoregressive rollouts [55]. Different approaches have been proposed to mitigate error accumulation. For instance, Pangu-Weather [56] employs a hierarchical temporal aggregation strategy, training four separate models with lead times of 1, 3, 6, and 24 hours. In contrast, FuXi adopts a cascade architecture, with models optimized for three sequential forecast intervals: 0–5 days (FuXi-Short), 5–10 days (FuXi-Medium), and 10–15 days (FuXi-Long).

In this study, we utilize FuXi-Short and FuXi-Medium. Notably, the transition from FuXi-Short to FuXi-Medium occurs on day 4, rather than on day 5 as originally proposed [33]. Supplementary Figs. 32 and 33 present std-based and RMSE-based activity, respectively, each normalized relative to the ECMWF HRES forecast activity. As a result, ECMWF HRES activity is scaled to 1 across the 10 days, and FuXi forecast activity values below 1 reflect reduced variability compared to ECMWF HRES. A distinct drop in forecast activity of FuXi Weather (red lines) is observed on day 4 (red lines), coinciding with the transition between the two FuXi models. Since our primary objective is to develop a cycling DA and forecasting system that incorporates satellite observations, improving forecast activity without compromising accuracy remains a key focus for future work. In this section, we present preliminary results demonstrating the effectiveness of an enhancement module applied to FuXi-Medium.

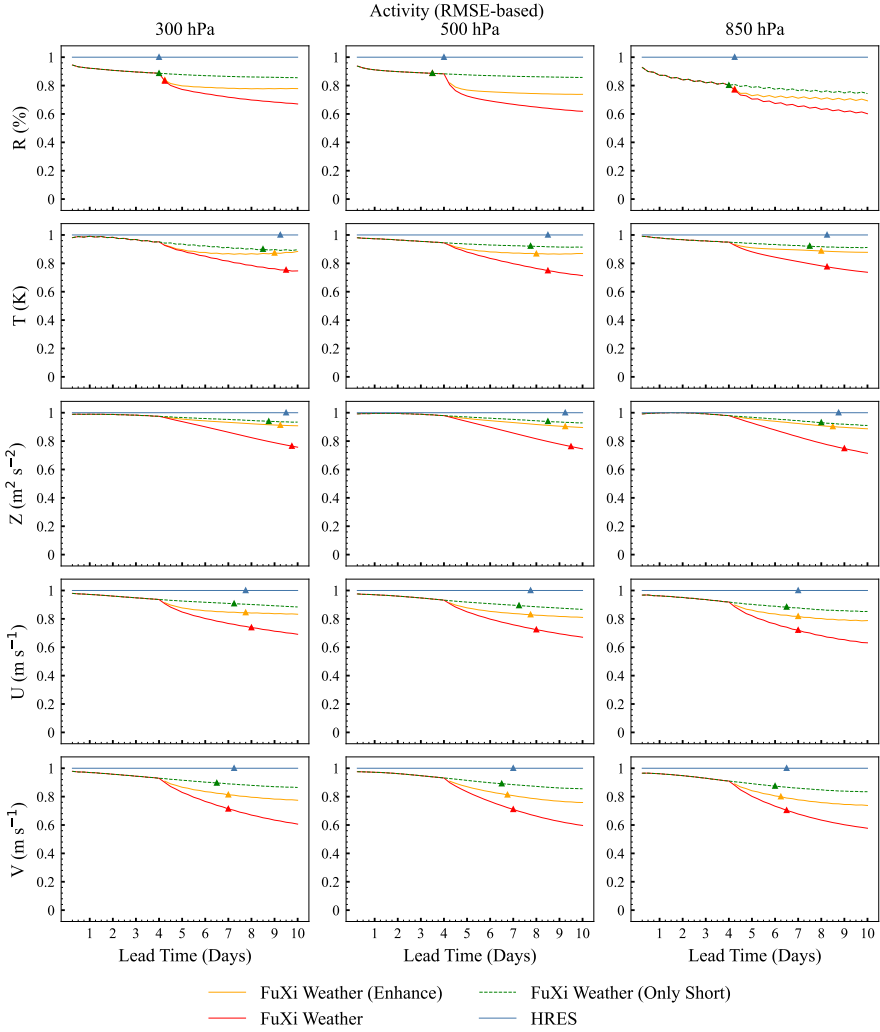
To mitigate forecast smoothness, we introduce an enhancement module appended to FuXi-Medium. This module consists of three repeated U-Net blocks, consistent with the architecture described in Section 2.2. The enhancement module is trained using a combined loss function comprising the structural similarity index (SSIM) [57] and L1 loss, weighted equally. SSIM measures perceptual similarity between images, ranging from 0 to 1 (where 1 indicates perfect similarity), and here serves as an indirect metric for forecast smoothness. During training, FuXi-Medium's parameters remain frozen, and only the enhancement module is updated. Due to GPU memory constraints, the number of autoregressive steps increases incrementally from 1 to 6. During inference, the enhancement module is applied after each FuXi-Medium prediction step. The optimized model is denoted as FuXi Weather (Enhance) in Supplementary Figs. 32 to 34, demonstrate significantly improved forecast activity compared to the original FuXi Weather while maintaining performance

comparable to the original FuXi Weather. Thus, the enhancement module successfully increases forecast activity with performance comparable to ECMWF HRES.

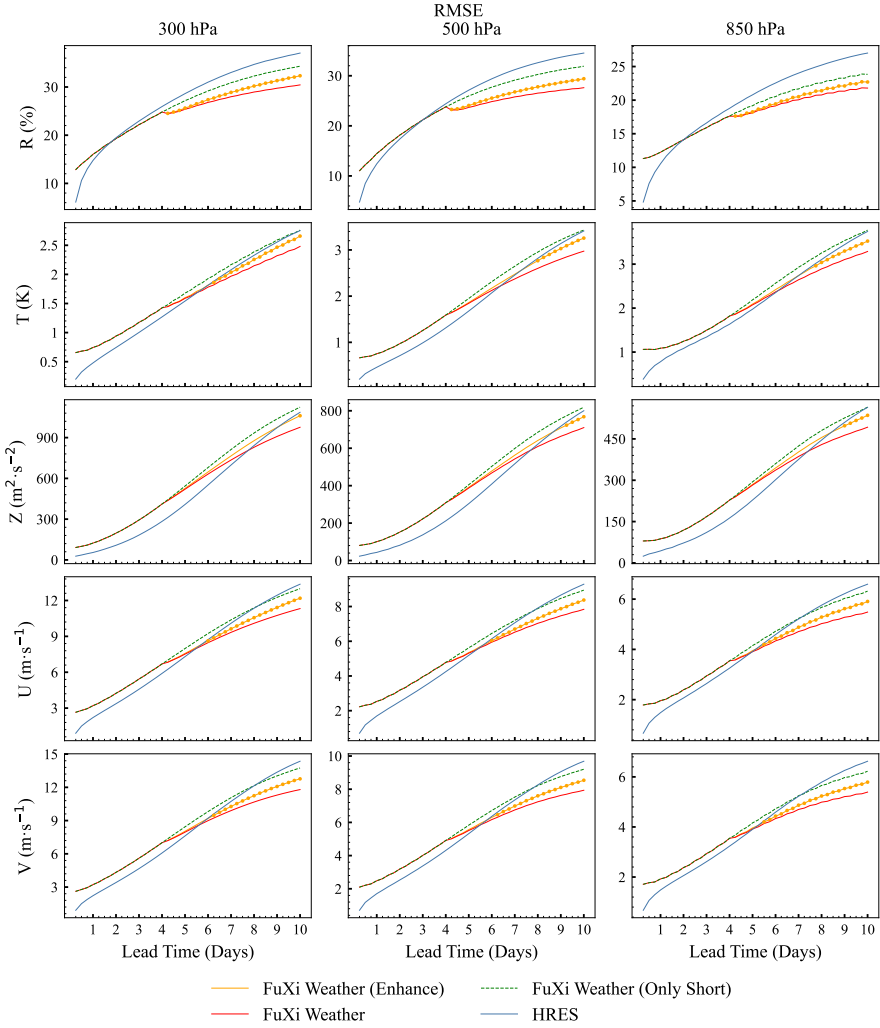
Moreover, recent studies suggest that generative models [58–60] and machine learning models integrating a differentiable solver for atmospheric dynamics [61, 62] mitigate the smoothing issue effectively. However, further investigation of these approaches falls outside the scope of this study.



Supplementary Figure 32: Comparison of std-based forecast activity among various models over a 1-year testing period, spanning July 03, 2023 - June 30, 2024. The figure presents the globally-averaged and latitude-weighted forecast activity for forecasts generated by the FuXi model and ECMWF HRES in 10-day forecasts. Forecast activity is scaled and centered relative to the ECMWF HRES activity. FuXi forecasts (red solid, green dashed, and yellow solid lines) are initialized using analysis fields from FuXi-DA. The green dashed lines represent FuXi-Weather (Short Only), consisting of 10-day forecasts produced solely by the FuXi-Short model, while the yellow lines denote FuXi-Weather (Enhance), where FuXi-Medium forecasts are refined by an enhancement module. The analysis includes five variables: relative humidity (RH), temperature (T), geopotential (Z), u component of wind (U), and v component of wind (V), at three pressure levels (300 hPa, 500 hPa, and 850 hPa). The five rows and three columns correspond to five variables and three pressure levels, respectively. Triangle markers superimposed on each line indicate skillful forecast lead times.



Supplementary Figure 33: Comparison of RMSE-based forecast activity among various models over a 1-year testing period, spanning July 03, 2023 - June 30, 2024. The figure presents the globally-averaged and latitude-weighted forecast activity for forecasts generated by the FuXi model and ECMWF HRES in 10-day forecasts. Forecast activity is scaled and centered relative to the ECMWF HRES activity. FuXi forecasts (red solid, green dashed, and yellow solid lines) are initialized using analysis fields from FuXi-DA. The green dashed lines represent FuXi-Weather (Short Only), consisting of 10-day forecasts produced solely by the FuXi-Short model, while the yellow lines denote FuXi-Weather (Enhance), where FuXi-Medium forecasts are refined by an enhancement module. The analysis includes five variables: relative humidity (RH), temperature (T), geopotential (Z), u component of wind (U), and v component of wind (V), at three pressure levels (300 hPa, 500 hPa, and 850 hPa). The five rows and three columns correspond to five variables and three pressure levels, respectively. Triangle markers superimposed on each line indicate skillful forecast lead times.



Supplementary Figure 34: Comparison of 10-day forecast performance across models over a 1-year testing period, spanning from July 03, 2023, to June 30, 2024. The figure presents the globally-averaged and latitude-weighted root mean square error (RMSE) for forecasts generated by the FuXi model and ECMWF HRES in 10-day forecasts. FuXi forecasts (red solid, green dashed, and yellow solid lines) are initialized using analysis fields from FuXi-DA. The green dashed lines represent FuXi-Weather (Short Only), consisting of 10-day forecasts produced solely by the FuXi-Short model, while the yellow lines denote FuXi-Weather (Enhance), where FuXi-Medium forecasts are refined by an enhancement module. The analysis includes five variables: relative humidity (RH), temperature (T), geopotential (Z), u component of wind (U), and v component of wind (V), at three pressure levels (300 hPa, 500 hPa, and 850 hPa). The five rows and three columns correspond to five variables and three pressure levels, respectively.

Supplementary References

- [1] Hersbach, H., *et al.*: The ERA5 global reanalysis. Q. J. R. Meteorol. Soc. **146**(730), 1999–2049 (2020). <https://doi.org/10.1002/qj.3803>
- [2] Eyre, J.R., English, S.J., Forsythe, M.: Assimilation of satellite data in numerical weather prediction. part i: The early years. Quarterly Journal of the Royal Meteorological Society **146**(726), 49–68 (2020)
- [3] Eyre, J.R., *et al.*: Assimilation of satellite data in numerical weather prediction. part ii: Recent years. Quarterly Journal of the Royal Meteorological Society **148**(743), 521–556 (2022). <https://doi.org/10.1002/qj.4228>
- [4] Joo, S., Eyre, J., Marriott, R.: The impact of metop and other satellite data within the met office global nwp system using an adjoint-based sensitivity method. Monthly Weather Review **141**(10), 3331–3342 (2013). <https://doi.org/10.1175/MWR-D-12-00232.1>
- [5] Goldberg, M.D., Kilcoyne, H., Cikanek, H., Mehta, A.: Joint polar satellite system: The united states next generation civilian polar-orbiting environmental satellite system. Journal of Geophysical Research: Atmospheres **118**(24), 13463–13475 (2013). <https://doi.org/10.1002/2013JD020389>
- [6] Zhou, L., Divakarla, M., Liu, X., Layns, A., Goldberg, M.: An overview of the science performances and calibration/validation of joint polar satellite system operational products. Remote Sensing **11**(6) (2019). <https://doi.org/10.3390/rs11060698>
- [7] Brotzge, J.A., *et al.*: Challenges and opportunities in numerical weather prediction. Bulletin of the American Meteorological Society **104**(3), 698–705 (2023). <https://doi.org/10.1175/BAMS-D-22-0172.1>
- [8] Harries, J.E.: Satellite meteorology: An introduction. by stanley q. kiddy and thomas h. von der haar. academic press. 1995. x + 466 pp. isbn 0 12 406430 2. Quarterly Journal of the Royal Meteorological Society **128**(581), 1037–1037 (2002). <https://doi.org/10.1002/qj.200212858118>
- [9] Jung, J.A., Zapotocny, T.H., Marshall, J.F.L., Treadon, R.E.: A two-season impact study of noaa polar-orbiting satellites in the ncep global data assimilation system. Weather and Forecasting **23**(5), 854–877 (2008). <https://doi.org/10.1175/2008WAF2007065.1>
- [10] Eyre, J.: The WMO Vision for global observing systems in 2025: to what extent will it be met by space agencies’ plans. In: Proceedings of the ECMWF Annual Seminar, Reading, UK, pp. 8–12 (2014).

<https://www.ecmwf.int/sites/default/files/elibrary/2015/9328-wmo-vision-global-observing-systems-2025-what-extent-will-it-be-met-space-agencies-plans.pdf>

- [11] Bormann, N., Lawrence, H., Farnan, J.: Global observing system experiments in the ECMWF assimilation system. European Centre for Medium Range Weather Forecasts (2019). <https://doi.org/10.21957/sr184iyz>. <https://www.ecmwf.int/en/elibrary/80953-global-observing-system-experiments-ecmwf-assimilation-system>
- [12] Zhang, P., et al.: Fy-3e: The first operational meteorological satellite mission in an early morning orbit. *Advances in Atmospheric Sciences* **39** (2022). <https://doi.org/10.1007/s00376-021-1304-7>
- [13] Shao, Y., et al.: System design and technical characteristics of fengyun-3e meteorological satellite. *Advances in Astronautics Science and Technology* **5**(4), 363–374 (2022)
- [14] Zhang, P., et al.: The on-orbit performance of fy-3e in an early morning orbit. *Bulletin of the American Meteorological Society* **105**(1), 144–175 (2024). <https://doi.org/10.1175/BAMS-D-22-0045.1>
- [15] Righetti, P., de Juana Gamon, J., Sancho, F.: Metop-c deployment and start of three-satellite operations. *The Aeronautical Journal* **124**(1276), 902–916 (2020). <https://doi.org/10.1017/aer.2020.10>
- [16] Asbury, S.C.: Jpss-1: An overview of the design, fabrication, test, launch, and early on-orbit performance of the nation's next-generation polar orbiting weather satellite. In: 2018 AIAA SPACE and Astronautics Forum and Exposition, pp. 1–16 (2018). <https://doi.org/10.2514/6.2018-5349>
- [17] Wang, Z., Román, M.O., Kalb, V.L., Miller, S.D., Zhang, J., Shrestha, R.M.: Quantifying uncertainties in nighttime light retrievals from suomi-npp and noaa-20 viirs day/night band data. *Remote Sensing of Environment* **263**, 112557 (2021). <https://doi.org/10.1016/j.rse.2021.112557>
- [18] Goldberg, M.: The joint polar satellite system overview. In: IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium, pp. 1581–1584 (2018). <https://doi.org/10.1109/IGARSS.2018.8518787>. IEEE
- [19] Kursinski, E.R., Hajj, G.A., Schofield, J.T., Linfield, R.P., Hardy, K.R.: Observing earth's atmosphere with radio occultation measurements using the global positioning system. *Journal of Geophysical Research: Atmospheres* **102**(D19), 23429–23465 (1997). <https://doi.org/10.1029/97JD01569>

- [20] Kursinski, E.R., Hajj, G.A., Leroy, S.S., Herman, B.: The gps radio occultation technique. *Terrestrial Atmospheric and Oceanic Sciences* **11**(1), 53–114 (2000). [https://doi.org/10.3319/TAO.2000.11.1.53\(COSMIC\)](https://doi.org/10.3319/TAO.2000.11.1.53(COSMIC))
- [21] Anthes, R.A.: Exploring earth’s atmosphere with radio occultation: contributions to weather, climate and space weather. *Atmospheric Measurement Techniques* **4**(6), 1077–1103 (2011). <https://doi.org/10.5194/amt-4-1077-2011>
- [22] Cardinali, C., Healy, S.: Impact of gps radio occultation measurements in the ecwrf system using adjoint-based diagnostics. *Quarterly Journal of the Royal Meteorological Society* **140**(684), 2315–2320 (2014). <https://doi.org/10.1002/qj.2300>
- [23] Bauer, P., Radnóti, G., Healy, S., Cardinali, C.: Gns radio occultation constellation observing system experiments. *Monthly Weather Review* **142**(2), 555–572 (2014). <https://doi.org/10.1175/MWR-D-13-00130.1>
- [24] Shehaj, E., Leroy, S., Cahoy, K., Geiger, A., Crocetti, L., Moeller, G., Soja, B., Rothacher, M.: Gns radio occultation climatologies mapped by machine learning and bayesian interpolation. *Atmospheric Measurement Techniques Discussions* **2023**, 1–24 (2023). <https://doi.org/10.5194/amt-2023-205>
- [25] Schreiner, W.S., Weiss, J., Anthes, R.A., Braun, J., Chu, V., Fong, J., Hunt, D., Kuo, Y.-H., Meehan, T., Serafino, W., et al.: Cosmic-2 radio occultation constellation: First results. *Geophysical Research Letters* **47**(4) (2020). <https://doi.org/10.1029/2019GL086841>
- [26] Anthes, R., Sjoberg, J., Feng, X., Syndergaard, S.: Comparison of cosmic and cosmic-2 radio occultation refractivity and bending angle uncertainties in august 2006 and 2021. *Atmosphere* **13**(5) (2022). <https://doi.org/10.3390/atmos13050790>
- [27] Foelsche, U., Kirchengast, G., Steiner, A.K., Kornblueh, L., Manzini, E., Bengtsson, L.: An observing system simulation experiment for climate monitoring with GNSS radio occultation data: Setup and test bed study. *Journal of Geophysical Research (Atmospheres)* **113**(D11), 1–14 (2008). <https://doi.org/10.1029/2007JD009231>
- [28] Guo, Y., Wang, H., Hu, Q., Liu, H., Liu, L., Bennamoun, M.: Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence* **43**(12), 4338–4364 (2020). <https://doi.org/10.1109/TPAMI.2020.3005434>
- [29] Bello, S.A., Yu, S., Wang, C., Adam, J.M., Li, J.: Deep learning on 3d point clouds. *Remote Sensing* **12**(11), 1729 (2020). <https://doi.org/10.3390/rs12111729>

3390/rs12111729

- [30] Zhou, Y., Tuzel, O.: Voxelnet: End-to-end learning for point cloud based 3d object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4490–4499 (2018). <https://doi.org/10.1109/CVPR.2018.00472>
- [31] Lang, A.H., *et al.*: PointPillars: Fast Encoders for Object Detection From Point Clouds . In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12689–12697. IEEE Computer Society, Los Alamitos, CA, USA (2019). <https://doi.org/10.1109/CVPR.2019.01298>. <https://doi.ieeecomputersociety.org/10.1109/CVPR.2019.01298>
- [32] Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer Normalization. Preprint at <https://arxiv.org/abs/1607.06450> (2016)
- [33] Chen, L., *et al.*: Fuxi: A cascade machine learning forecasting system for 15-day global weather forecast. *npj Climate and Atmospheric Science* **6**(190), 1–11 (2023). <https://doi.org/10.1038/s41612-023-00512-1>
- [34] Harris, B., Kelly, G.: A satellite radiance-bias correction scheme for data assimilation. *Q. J. R. Meteorol. Soc.* **127**(574), 1453–1468 (2001). <https://doi.org/10.1002/qj.49712757418>
- [35] Auligné, T., McNally, A.P., Dee, D.P.: Adaptive bias correction for satellite data in a numerical weather prediction system. *Q. J. R. Meteorol. Soc.* **133**(624), 631–642 (2007). <https://doi.org/10.1002/qj.56>
- [36] Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., Sun, C.: Attention bottlenecks for multimodal fusion. *Advances in neural information processing systems* **34**, 14200–14213 (2021). <https://doi.org/10.5555/3540261.3541349>
- [37] Eyre, J.: A bias correction scheme for simulated TOVS brightness temperatures. ECMWF Tech. Memo. 186. (1992). <https://doi.org/10.21957/tmhrqv5cp>
- [38] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 234–241 (2015). https://doi.org/10.1007/978-3-319-24574-4_28. Springer
- [39] Elfwing, S., Uchibe, E., Doya, K.: Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks* **107**, 3–11 (2018). <https://doi.org/10.1016/j.neunet.2017.12.012>
- [40] Shi, W., *et al.*: Real-time single image and video super-resolution using an

efficient sub-pixel convolutional neural network. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1874–1883 (2016).
<https://doi.org/10.1109/cvpr.2016.207>

[41] Paszke, A., *et al.*: Automatic differentiation in pytorch. In: NIPS 2017 Workshop on Autodiff (2017).
<https://openreview.net/forum?id=BJJsrnfCZ>

[42] Loshchilov, I., Hutter, F.: SGDR: Stochastic gradient descent with warm restarts. In: International Conference on Learning Representations (2017).
<https://openreview.net/forum?id=Skq89Scxx>

[43] Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning. In: International Conference on Learning Representations (2017).
<https://openreview.net/forum?id=rJY0-Kcll>

[44] Wang, Y., Yao, Q., Kwok, J.T., Ni, L.M.: Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv.* **53**(3) (2020).
<https://doi.org/10.1145/3386252>

[45] Van de Ven, G.M., Tuytelaars, T., Tolias, A.S.: Three types of incremental learning. *Nature Machine Intelligence* **4**(12), 1185–1197 (2022). <https://doi.org/10.1038/s42256-022-00568-3>

[46] Lam, R., *et al.*: Learning skillful medium-range global weather forecasting. *Science* **382**(6677), 1416–1421 (2023). <https://doi.org/10.1126/science.adi2336>

[47] Hocking, J., Vidot, J., Brunel, P., Roquet, P., Silveira, B., Turner, E., Lupu, C.: A new gas absorption optical depth parameterisation for rrtov version 13. *Geoscientific Model Development* **14**(5), 2899–2915 (2021).
<https://doi.org/10.5194/gmd-14-2899-2021>

[48] Xu, X., Sun, X., Han, W., Zhong, X., Chen, L., Gao, Z., Li, H.: Fuxi-da: A generalized deep learning data assimilation framework for assimilating satellite observations. *npj Climate and Atmospheric Science* **8**(156), 1–15 (2025). <https://doi.org/10.1038/s41612-025-01039-3>

[49] Allen, A., Markou, S., Tebbutt, W., Requeima, J., Bruinsma, W.P., Andersson, T.R., Herzog, M., Lane, N.D., Chantry, M., Hosking, J.S., *et al.*: End-to-end data-driven weather prediction. *Nature*, 1–30 (2025).
<https://doi.org/10.1038/s41586-025-08897-0>

[50] Wang, B., Zou, X., Zhu, J.: Data assimilation and its applications. *Proceedings of the National Academy of Sciences* **97**(21), 11143–11144 (2000).
<https://doi.org/10.1073/pnas.97.21.11143>

- [51] Zhang, K., *et al.*: A methodology to adjust atms observations for limb effect and its applications. *Journal of Geophysical Research: Atmospheres* **122**(21), 11347–11356 (2017). <https://doi.org/10.1002/2017JD026820>
- [52] Li, X., Zou, X., Zeng, M.: An alternative bias correction scheme for cris data assimilation in a regional model. *Monthly Weather Review* **147**(3), 809–839 (2019). <https://doi.org/10.1175/MWR-D-18-0044.1>
- [53] Weng, F., Zou, X., Wang, X., Yang, S., Goldberg, M.: Introduction to suomi national polar-orbiting partnership advanced technology microwave sounder for numerical weather prediction and tropical cyclone applications. *Journal of Geophysical Research: Atmospheres* **117**(D19) (2012). <https://doi.org/10.1029/2012JD018144>
- [54] Hoffman, R.N., Liu, Z., Louis, J.-F., Grassoti, C.: Distortion representation of forecast errors. *Monthly Weather Review* **123**(9) (1995). [https://doi.org/10.1175/1520-0493\(1995\)123<2758:DROFE>2.0.CO;2](https://doi.org/10.1175/1520-0493(1995)123<2758:DROFE>2.0.CO;2)
- [55] Siddiqui, S.A., *et al.*: Exploring the design space of deep-learning-based weather forecasting systems. Preprint at <https://arxiv.org/abs/2410.07472> (2024)
- [56] Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., Tian, Q.: Accurate medium-range global weather forecasting with 3d neural networks. *Nature* **619**(7970), 533–538 (2023). <https://doi.org/10.1038/s41586-023-06185-3>
- [57] Brunet, D., Vrsnay, E.R., Wang, Z.: On the mathematical properties of the structural similarity index. *IEEE Transactions on Image Processing* **21**(4), 1488–1499 (2012). <https://doi.org/10.1109/TIP.2011.2173206>
- [58] Zhong, X., *et al.*: FuXi-Extreme: Improving extreme rainfall and wind forecasts with diffusion model. *Sci. China Earth Sci.* **67**, 3696–3708 (2024). <https://doi.org/10.1007/s11430-023-1427-x>
- [59] Couairon, G., Singh, R., Charantonis, A., Lessig, C., Monteleoni, C.: ArchesWeather & ArchesWeatherGen: a deterministic and generative model for efficient ML weather forecasting. Preprint at <https://arxiv.org/abs/2412.12971> (2024)
- [60] Price, I., *et al.*: Probabilistic weather forecasting with machine learning. *Nature* **637**(8044), 84–90 (2025). <https://doi.org/10.1038/s41586-024-08252-9>
- [61] Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., Klöwer, M., Lottes, J., Rasp, S., Düben, P., *et al.*: Neural general circulation models for weather and climate. *Nature* **632**(8027), 1060–1066 (2024). <https://doi.org/10.1038/s41586-024-07744-y>

- 1025 [62] Verma, Y., Heinonen, M., Garg, V.: ClimODE: Climate and
1026 weather forecasting with physics-informed neural ODEs. In: The
1027 Twelfth International Conference on Learning Representations (2024).
1028 <https://openreview.net/forum?id=xuY33XhEGR>