ATMOSPHERIC SCIENCE

# FuXi-ENS: A machine learning model for efficient and accurate ensemble weather prediction

**Xiaohui Zhong[1]†, Lei Chen[1,2]†, Hao Li[1,2]*†, Roberto Buizza[3], Jun Liu[1], Jie Feng[4]*, Zijian Zhu[1], Xu Fan[2], Kan Dai[5], Jing-jia Luo[6], Jie Wu[7], Bo Lu[7,8]***



**Ensemble forecasting is essential for quantifying forecast uncertainty and providing probabilistic weather predictions. However, the substantial computational demands of current global ensemble prediction systems based on conventional models limit ensemble sizes, hindering the representation of diverse weather scenarios. Recent advances in machine learning (ML) have greatly reduced computational costs and improved deterministic forecasting. Nonetheless, applying ML to ensemble forecasting poses challenges in addressing uncertainties in initial conditions and models, which are the major sources of forecasting errors. To address these challenges, we introduce FuXi-ENS, an advanced ML model that generates 6-hourly global ensemble weather forecasts up to 15 days ahead at a spatial resolution of 0.25°. Using a variational autoencoder framework, FuXi-ENS optimizes a loss function that combines the continuous ranked probability score (CRPS) with the Kullback-Leibler divergence, enabling flow-dependent perturbations. Comprehensive evaluations demonstrate that FuXi-ENS outperforms the ECMWF ensemble in key forecast metrics such as CRPS and Brier score.**

## INTRODUCTION

Weather forecasting is inherently uncertain due to the chaotic and nonlinear nature of the atmosphere (*1*). Quantifying and reducing these uncertainties is crucial for improving forecast accuracy and reliability (*2*). Uncertainty estimates are particularly valuable (*3*, *4*) in weather-sensitive sectors such as risk assessment (*5*, *6*), renewable energy (*7*, *8*), and aviation, where reliable forecasts enhance safety and operational efficiency (*9*). Major sources of uncertainty include imperfections in forecast models and errors in initial conditions.

Ensemble forecasting, which involves running multiple simulations with slight variations in initial conditions and model physics (*10*, *11*), provides today the only feasible way to estimate forecast uncertainty through the spread among individual members. These diverse simulations allow for probabilistic forecasts by predicting the likelihood of various outcomes. Traditional ensemble prediction systems (EPSs) are based on the physics-based numerical weather prediction (NWP) models. For example, the state-of-the-art EPS of European Centre for Medium-Range Weather Forecasts (ECMWF; ECMWF-ENS hereafter) uses techniques like the singular vector (SV) approach (*12*–*14*) and the stochastically perturbed parameterization tendency scheme (*15*, *16*) to introduce perturbations in initial conditions and physical tendencies, respectively, within its deterministic Integrated Forecast System (IFS). While increasing the ensemble size can improve uncertainty estimates (*15*, *17*), the substantial computational demands of NWP models constrain ensemble size, limiting the ability to represent the full probability distribution of possible weather scenarios. Major global ensemble systems typically have between 14 and 51 members (*18*–*20*), underscoring the need for more computationally efficient approaches to ensemble forecasting.

Recent advancements in machine learning (ML) have substantially improved the computational efficiency and accuracy of global weather forecasts, providing a promising alternative to traditional NWP methods (*21*–*25*). While early ML applications focused on deterministic forecasting, recent efforts have begun to explore the more complex and challenging domain of ensemble forecasting. For instance, Chen *et al.* (*24*) used random Perlin noise to perturb initial conditions, but this approach led to a reduced ensemble spread at longer lead times, resulting in an underestimation of forecast uncertainty. Beyond initial-condition perturbations, generative ML models have been explored as potential solutions. For example, Price *et al.* (*26*) developed GenCast using a diffusion model (*27*, *28*) to account for uncertainties in initial and subsequent predicted weather states. Li *et al.* (*29*) proposed the Scalable Ensemble Envelope Diffusion Sampler (SEEDS), which considers the spatial coherence of variables. However, SEEDS requires input from two Global Ensemble Forecast System (GEFS) (*20*) members and is limited to predicting only eight variables at a coarse 2° resolution. Despite these advances (*24*, *26*, *29*–*31*), an optimal ML-based ensemble forecasting method remains elusive, particularly for predicting extreme weather events at fine spatial resolutions.

In this study, we introduce FuXi-ENS, an ML-based ensemble model for medium-range weather forecasting at a spatial resolution of 0.25°. Unlike existing ML ensemble models such as SEEDS, which depend on external ensemble data like GEFS, FuXi-ENS adapts the traditional ensemble forecasting framework to an ML model. It introduces explainable perturbations in both the initial conditions and forecast steps by training on the fifth generation ECMWF reanalysis (ERA5) datasets. These perturbations are generated by a perturbation model optimized with an innovative loss function uniquely combining the continuous ranked probability score (CRPS) with Kullback-Leibler (KL) divergence. The CRPS, a well-established metric for probabilistic

[1]Artificial Intelligence Innovation and Incubation Institute, Fudan University, Shanghai, China. [2]Shanghai Academy of Artificial Intelligence for Science, Shanghai, China. [3]Scuola Universitaria Superiore Sant'Anna, Pisa, Italy. [4]Department of Atmospheric and Oceanic Sciences and Institute of Atmospheric Sciences, Fudan University, Shanghai, China. [5]National Meteorological Information Center, China Meteorological Administration, Beijing, China. [6]Institute for Climate and Application Research (ICAR)/CIC-FEMD/KLME/ILCEC, Nanjing University of Information Science and Technology, Nanjing, China. [7]State Key Laboratory of Climate System Prediction and Risk Management/China Meteorological Administration Climate Studies Key Laboratory National Climate Center, Beijing, China. [8]Xiong'an Institute of Meteorological Artificial Intelligence, Xiong'an New Area, China.
*Corresponding author. Email: lihao_lh@fudan.edu.cn (H.L.); fengjiefj@fudan.edu.cn (J.F.); bolu@cma.gov.cn (B.L.)
†These authors contributed equally to this work.

forecasting, imposes a constraint on the probability distribution of forecast states. In contrast to the *L1* loss, which optimizes for the mean of the distribution and is suited for deterministic forecasting, CRPS is a proper scoring rule for assessing probabilistic forecasts by comparing the forecast's cumulative distribution function with the observed value. This makes FuXi-ENS better suited for ensemble forecast applications compared to traditional variational autoencoder (VAE) models, which typically use *L1* loss combined with KL divergence.

As a result of this innovative scheme, FuXi-ENS outperforms ECMWF-ENS in several key deterministic and probability verification metrics, such as the root mean square error (RMSE), anomaly correlation coefficient (ACC), CRPS, mean bias error (MBE), Brier score (BS), rank histogram, and receiver operating characteristic area (ROCA) skill score (ROCASS). FuXi-ENS also demonstrates superior performance in forecasting extreme weather events, such as tropical cyclones (TCs) and heatwaves, compared to ECMWF-ENS. In particular, FuXi-ENS surpasses traditional NWP ensembles and diffusion-based ML models in computational efficiency, completing a 15-day forecast with 6-hour intervals in just 10 s per member on an Nvidia A100 graphics processing unit (GPU). The model's accuracy and efficiency have the potential to reduce false alarms and enhance emergency preparedness, which are critical for mitigating threats to life and property (*32*, *33*).

## RESULTS

This study presents a comprehensive evaluation of FuXi-ENS global forecasts over a 1-year testing period in 2018, using eight A100 GPUs to generate 48 ensemble members (six members per GPU). The performance of the 48-member FuXi-ENS is compared to the 51-member ECMWF-ENS. To ensure a thorough comparison, various verification metrics (*34*) were used, including deterministic skill scores for ensemble mean forecasts and probabilistic metrics derived from all ensemble members, such as the spread-skill ratio (SSR) (*35*), CRPS, BS, ROCASS curves, and rank histogram. The evaluation also focused on the prediction of extreme weather events, particularly TC tracks and the record-breaking 2018 Northeast Asia heatwave. As supplementary analyses, the energy spectra of prognostic variables were examined for both ensemble mean and individual members. Additional evaluations, including spread-RMSE correlation analyses, assessments of typical atmospheric balance dynamics such as the gradient wind and hydrostatic balance, and forecast comparisons between FuXi-ENS (48 members) and GenCast (50 members) are presented in the Supplementary Materials. Comprehensive descriptions of all evaluation methodologies are provided in the Supplementary Materials.

### Deterministic forecast

We first evaluate the performance of ensemble mean forecasts using standard verification metrics for deterministic forecasts, including RMSE, ACC, and MBE (*36*, *37*). RMSE quantifies the average magnitude of forecast errors, serving as a direct measure of accuracy, while MBE measures the systematic deviations between long-term forecast mean and the observed reference. Both RMSE and MBE are negatively oriented, with values closer to zero indicating better performance. ACC, a positively oriented metric, measures the correlation between predicted and observed anomalies, reflecting the model's capability to capture large-scale synoptic patterns.

As shown in Fig. 1, FuXi-ENS consistently shows lower RMSE than ECMWF-ENS across most lead times and variables, with statistically significant improvements, especially within the first 7 days. The improvement is most notable for 2-m temperature (T2M), where FuXi-ENS reduces RMSE by up to 25% around day 2, compared to ~10% for other variables. This superior performance in T2M may be attributed to notably lower forecast bias in FuXi-ENS compared to ECMWF-ENS. Notice that T2M is a diagnostic variable in ECMWF IFS, where biases may arise from errors associated with factors such as topography. In contrast, FuXi-ENS predicts T2M using a training dataset that includes reanalysis data for T2M and related variables, potentially reducing forecast bias by avoiding the limitations inherent in the diagnostic approach. The forecast bias for WS850 in FuXi-ENS is smaller than that in ECMWF-ENS during the first 5 days but becomes larger thereafter. For another prognostic variable Z500, FuXi-ENS exhibits a positive bias roughly double the negative bias observed in ECMWF-ENS, possibly linked to larger biases of FuXi-ENS in other variables, such as T850 and MSL, which are used to diagnose Z500. Consistent with RMSE, FuXi-ENS demonstrates higher ACC scores than ECMWF-ENS. Additional atmospheric variables at different pressure levels show similar trends in RMSE and ACC (see fig. S1).

In addition to the globally averaged error, figs. S2 and S3 provide grid-point comparisons of sample-mean RMSE (without latitude weighting) at forecast days 5, 10, and 15 for ensemble mean forecasts from ECMWF-ENS and FuXi-ENS. At day 5, FuXi-ENS shows lower RMSE over ~70% of grid points for six key variables, aligning with its improved spatial-mean RMSE and ACC performance. As lead times extend to 10 and 15 days, the performance of ECMWF-ENS and FuXi-ENS becomes more comparable, with RMSE differences varying by region.

### Probabilistic forecast

Unlike deterministic forecasts, ensemble forecasts offer the distinct advantage of providing probabilistic guidance. To evaluate and compare the probabilistic forecast skill of ECMWF-ENS and FuXi-ENS, we use several standard metrics, including CRPS, ROCASS (*38*), SSR, and BS scores. CRPS and BS measure the alignment between the ensemble's predicted probability distribution and the observed outcomes, providing insights into forecast reliability and resolution (discrimination skill) (*39*) (see section S1 for more details). SSR evaluates forecast reliability by comparing ensemble spread (i.e., the variability among ensemble members) with the ensemble mean error, where an ideal SSR value is close to one, signifying that the spread effectively represents forecast uncertainty by matching the magnitude of ensemble mean error.

Figure 2 illustrates the temporal evolution of globally averaged and latitude-weighted CRPS, ROCASS, and SSR for the same three variables shown in Fig. 1, over all 15-day forecast lead times. Results for additional variables, including T850, MSL, and WS10M, which are qualitatively similar, are shown in fig. S4. Overall, FuXi-ENS outperforms ECMWF-ENS in CRPS for nearly all variables and lead times, except for Z500 during the first 2 days and T850 within the initial 12 hours. Notably, FuXi-ENS shows the greatest CRPS improvement within the first 7 days, with relative improvements of 10 to 30%, which are statistically significant. This suggests the effectiveness of FuXi-ENS's probability-based loss function in enhancing ensemble sampling. However, the improvement diminishes as lead time increases, likely due to growing nonlinear effects in predicting

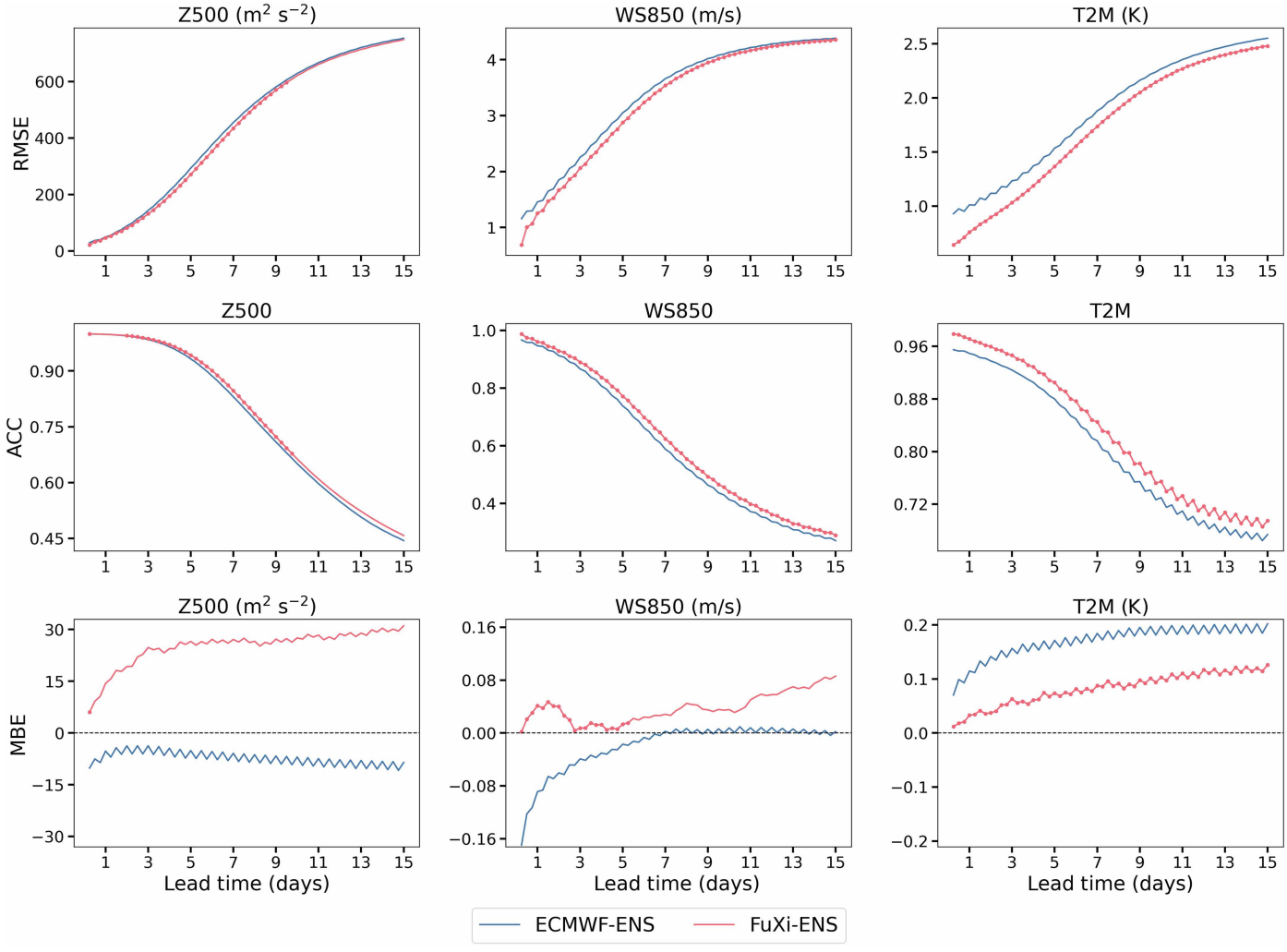


**Fig. 1. Comparison of ensemble mean forecast performance.** Comparison of globally averaged and latitude-weighted RMSE (first row), ACC (second row), and MBE (third row) of ensemble mean forecasts from ECMWF-ENS (blue lines) and FuXi-ENS (red lines). The figure includes three variables, such as 500-hPa geopotential (Z500, first column), 850-hPa wind speed (WS850, second column), and 2-m temperature (T2M; third column), in 15-day forecasts using testing data from 2018. A bootstrapping method, repeated 1000 times, was applied for significance testing. Red dots indicate where FuXi-ENS shows statistically significant improvement over ECMWF-ENS at the 97.5% confidence level.

the historical probability of prognostic variables during the training process of FuXi-ENS. In addition to CRPS, the ROCASS metric also presents overall better performance for FuXi-ENS compared to ECMWF-ENS, with improvements statistically significant for most lead times, except for T2M. This indicates that FuXi-ENS offers better performance in resolution for both lower and upper terciles, with an enhanced detection rate and reduced false alarms.

We also compare globally averaged SSR between ECMWF-ENS and FuXi-ENS as a function of lead times. For five of the six variables evaluated (Z500 and T2M in Fig. 2 and T850, MSL, and WS10M in fig. S4), FuXi-ENS initially exhibits SSR values closer to one, suggesting a closer alignment between ensemble spread and the mean square error of the ensemble mean compared to ECMWF-ENS. However, FuXi-ENS exhibits notable underestimation of ensemble spread, by 20 to 30%, around day 2, while ECMWF-ENS maintains near-optimal SSR values. This underestimation in FuXi-ENS may result from the limited projection of initial perturbations onto optimally growing modes, restricting the amplification of forecast perturbations. In contrast, ECMWF-ENS uses SVs as initial perturbations, which demonstrate optimal linear growth at early lead times. By appropriately scaling these perturbations to reflect analysis error statistics and further enhancing them through precisely calibrated stochastic model perturbations, the divergence of the ensemble spread is substantially enhanced. Beyond day 2, FuXi-ENS shows a marked increase in ensemble spread, reaching over 90% of the ensemble mean error around day 7. This trend may result from continuous perturbations constrained by the CRPS-related probability distribution applied at each forecasting step. At short lead times, when the forecast uncertainty is relatively low, the CRPS closely resembles the *L1* loss, resulting in less pronounced perturbations during these initial stages. As lead time increases, increasing uncertainty amplifies these perturbations. Further exploration into incorporating optimally growing perturbations during model training could improve FuXi-ENS performance.

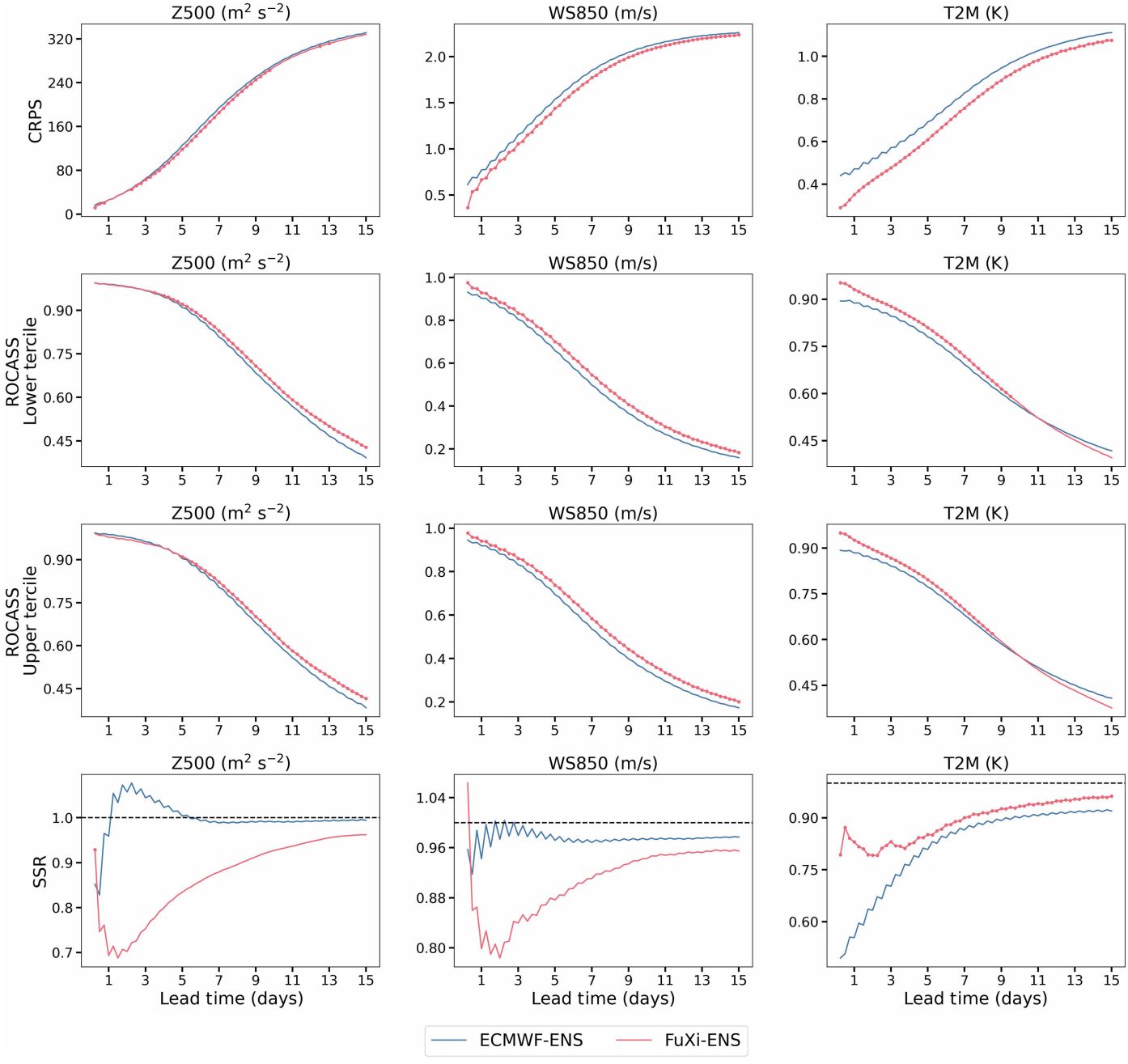


**Fig. 2. Comparison of probabilistic forecast performance.** Comparison of globally averaged and latitude-weighted CRPS (first row), ROCASS for lower tercile (second row) and upper tercile (third row), and SSR (fourth row) of ensemble forecasts from ECMWF-ENS (blue lines) and FuXi-ENS (red lines). The figure includes three variables, such as 500-hPa geopotential (Z500, first column), 850-hPa wind speed (WS850, second column), and T2M (third column), in 15-day forecasts using testing data from 2018. A bootstrapping method, repeated 1000 times, was applied for significance testing. Red dots indicate where FuXi-ENS shows statistically significant improvement over ECMWF-ENS at the 97.5% confidence level.

Similar to figs. S2 and S3, figs. S5 and S6 depict the spatial distribution of CRPS for ECMWF-ENS and FuXi-ENS. FuXi-ENS generally outperforms ECMWF-ENS in grid-point CRPS for four of the six variables, including WS850, T850, WS10M, and T2M, as indicated by the prevalence of blue and white areas. Notably, Z500 shows the greatest variability in CRPS difference. These patterns suggest that FuXi-ENS provides more accurate probabilistic forecasts in terms of CRPS than ECMWF-ENS, particularly at shorter lead times and for near-surface variables like T2M and WS10M variables. Additionally, figs. S7 to S10 present receiver operating characteristic (ROC) curves for ensemble forecasts from ECMWF-ENS and FuXi-ENS at lead times of 5, 10, and 15 days, comparing six variables in both the upper and lower terciles. The results indicate that FuXi-ENS demonstrates overall better probabilistic forecast skill than ECMWF-ENS, and

both models show diminished performance with increasing lead time, consistent with the ROCASS comparisons shown in Fig. 2 and fig. S4.

### Extreme forecast

The capability of an EPS to capture possible extreme events, which lie in the tails of the probability distribution, is a key for assessing ensemble forecast reliability and resolution. In this study, we assess the performance of ECMWF-ENS and FuXi-ENS in predicting extreme events using the BS (*40*) across different categories. The BS measures the accuracy of probabilistic forecasts for outcomes that exceed or fall below specific percentiles, based on the statistical probability distribution for each variable. These percentiles are derived from 24 years of ERA5 data (from 1993 to 2016), varying by grid point, month, and time of day.

Figure 3 compares the globally averaged and latitude-weighted BS between ECMWF-ENS and FuXi-ENS for events exceeding the 95th percentiles, as well as those below the 5th percentiles. These percentiles represent extreme high and low events for Z500, WS850, and T2M across all lead times in 15-day forecasts. Similar trends are observed for additional percentiles (above 90 and 98% and below 10 and 2%) and variables (see fig. S11). This analysis includes 2160 cases, derived from 6 percentiles, 6 variables, and 60 forecast lead times ($2160 = 6 \times 6 \times 60$). Consistent with CRPS performance, the BS improvement for FuXi-ENS is more pronounced at shorter lead times. The following subsections provide more detailed evaluations of the probabilistic TC track forecasts and predictions for the record-breaking 2018 Northeast Asia heatwave.

### Probabilistic TC track forecast

Accurate prediction of TC tracks is crucial for public safety and economic stability, given the severe impact of high winds, heavy rainfall, and storm surges during TC landfalls (*41*). Over past two decades, TC track forecast accuracy has improved substantially, with errors reduced by approximately two-thirds. However, some studies [e.g., (*42*)] suggest that the predictive skill of TC tracks may be approaching its limits, as indicated by trends in forecast errors for the North Atlantic and eastern North Pacific regions. Despite this, recent ML-based deterministic weather forecasting models (*22*, *23*, *43*) have demonstrated improved TC track prediction accuracy compared to ECMWF high-resolution forecast, suggesting potential for further improvements. In practice, end users require not only a single track prediction but also reliable uncertainty estimates. Traditional deterministic models (*44*) are inadequate for quantifying this uncertainty. In contrast, ensemble forecasts provide scenario-dependent uncertainty estimates (*45*), offering a more comprehensive understanding of possible outcomes.

Figure 4 presents box plots of the time-accumulated mean TC position error ($\mathrm{AccERROR_{TC}}$) and ensemble spread ($\mathrm{AccSpread_{TC}}$) for FuXi-ENS and ECMWF-ENS across all forecast cases, as a function of lead time. It also shows the forecast biases for TC position errors, decomposed into along-track ($\mathrm{AT_{TC}}$) and cross-track ($\mathrm{CT_{TC}}$) components. The assessment includes 167 predictions from FuXi-ENS and 165 from ECMWF-ENS, covering 20% of the 2018 testing data, using the International Best Track Archive for Climate Stewardship (IBTrACS) as a benchmark. Forecasts were only included if at least two-thirds of ensemble members (i.e., 32 members) detected TCs (*45*).



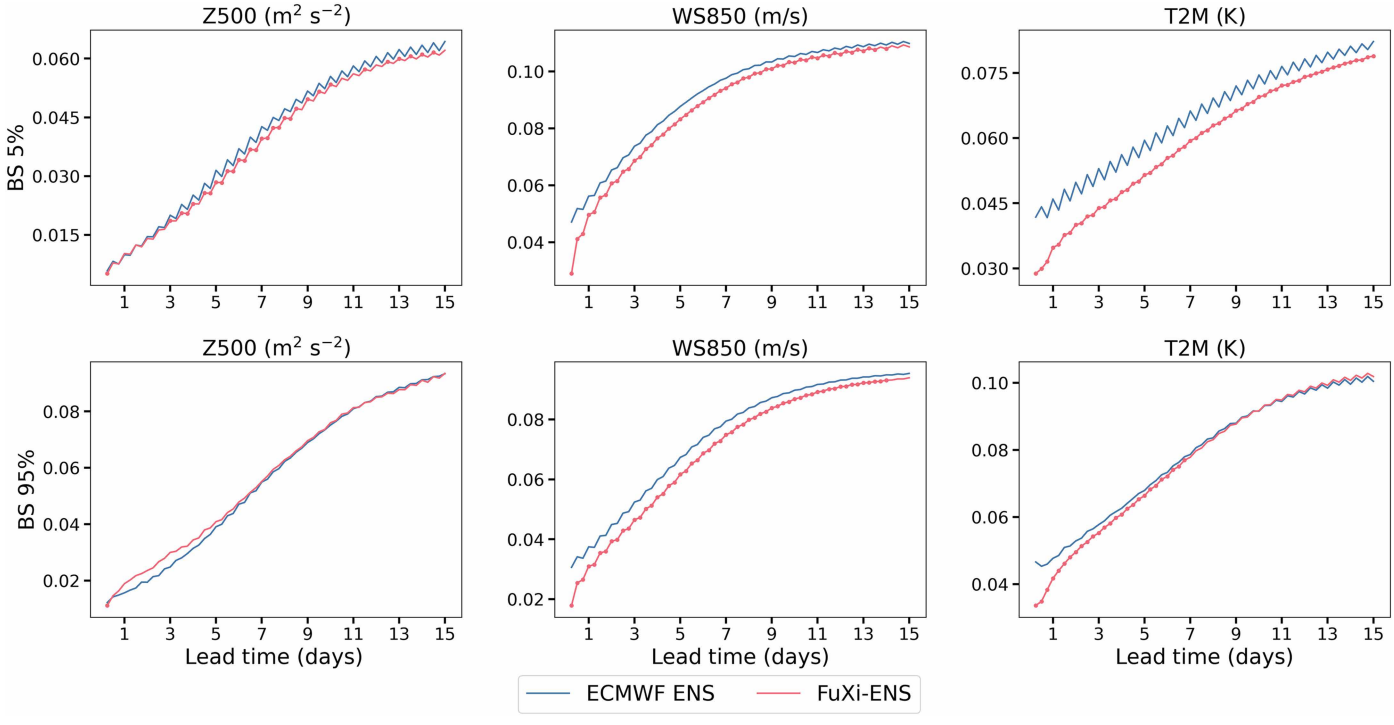


**Fig. 3. Comparison of extreme forecast performance.** Comparison of ECMWF-ENS (blue lines) and FuXi-ENS (red lines) ensembles on globally averaged and latitude-weighted BS for <5th (first row) and >95th (second row) percentile events. The figure includes three variables, such as 500-hPa geopotential (Z500), 850-hPa wind speed (WS850), and T2M, in 15-day forecasts using testing data from 2018. A bootstrapping method, repeated 1000 times, was applied for significance testing. Red dots indicate where FuXi-ENS shows statistically significant improvement over ECMWF-ENS at the 97.5% confidence level.

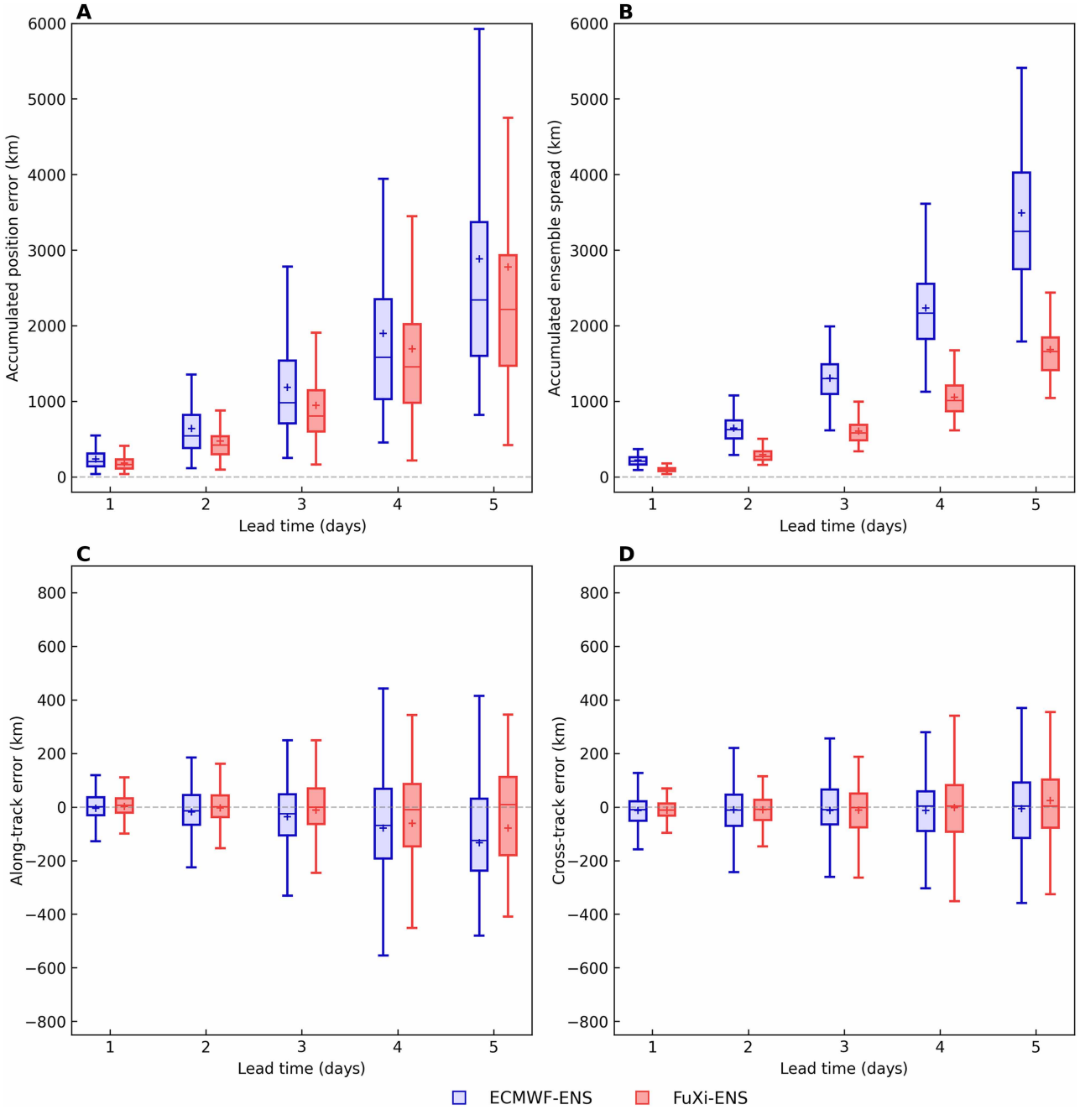


**Fig. 4. Comparison of TC forecast performance.** Box plot of TC forecast errors in 5-day forecasts for ECMWF-ENS (blue) and FuXi-ENS (red). (**A**) The accumulated mean position error ($AccERROR_{TC}$). (**A**) The accumulated ensemble spread ($AccSpread_{TC}$). (**C**) The along-track ($AT_{TC}$). (**D**) The cross-track ($CT_{TC}$). The 25th, 50th, and 75th percentiles are illustrated, together with the median value shown as + sign.

Overall, FuXi-ENS consistently shows smaller ensemble mean TC track errors than ECMWF-ENS at all lead times, with improvements ranging from 10% at early times to 20% by day 5. In most individual cases (see fig. S12), FuXi-ENS also exhibits lower TC track errors. Although FuXi-ENS provides more skillful ensemble mean forecasts, it tends to underestimate uncertainty, especially at longer lead times, with underestimations of 20 to 30% beyond day 3. A scatter plot of ensemble mean track error versus ensemble spread for FuXi-ENS (fig. S12) reveals that 118 forecast cases lie above the diagonal at day 3, compared to 49 below. In contrast, ECMWF-ENS overestimates uncertainty by about 15%, with 56 cases above and 109 below the diagonal. These differences in uncertainty estimation are likely related to the systems' respective SSR performance for synoptic-scale wind circulation (Fig. 2).

Moreover, along-track and cross-track biases reflect differences in TC speed and direction relative to observations, indicating whether the foretasted TC moves slower or faster or shifts to left or right. The results indicate that, while both FuXi-ENS and ECMWF-ENS exhibit

negligible cross-track bias, both demonstrate a slight slower along-track bias. However, FuXi-ENS has a smaller along-track bias, ~30 km less than ECMWF-ENS. Further analysis of individual cases shows that at day 5, FuXi-ENS forecast positions are more evenly distributed across the four quadrants surrounding the observed TC position, while ECMWF-ENS has a notable bias, with 97 of the 168 cases falling into the lower quadrant (not shown).

### Prediction of the record-breaking Northeast Asia heatwave in 2018

Heatwaves, characterized by prolonged periods of extreme heat, have had increasingly severe impacts on human health, ecosystems, agriculture, and infrastructure, due to global warming. During the summer of 2018, Northeast Asia experienced a record-breaking heatwave, which affected Japan, the Korean Peninsula, and northeastern China (*46*, *47*), resulting in substantial economic damage and numerous fatalities. Notably, Japan recorded its highest-ever temperature of 41.1°C in Kumagaya on July 23, while Seoul, South Korea, experienced its hottest day in 111 years at 39.6°C on 1 August. Northeastern China also saw temperatures nearing 39°C in Liaoning and Jilin provinces throughout July and August 2018. The heatwave caused at least 138 deaths in Japan and 42 in South Korea, with more than 7000 hospitalizations recorded. These tragic outcomes highlight the urgent need for accurate extreme weather forecasts to issue timely warnings and mitigate associated risks.

Figure 5 compares the 5-day ensemble forecast performance of T2M between FuXi-ENS and ECMWF-ENS during the 2018 Northeast Asia heatwave. ERA5 reanalysis data, shown in the first column, serve as the benchmark, depicting the spatial distribution of T2M at 6:00 UTC on 23 July 2018. The subsequent columns display 5-day forecasts from both ensembles, including the best-performing member, ensemble mean, and ensemble spread, all valid at the same time. The rankings of different ensemble members are based on the T2M RMSE averaged over the depicted land area. Results indicate that the best

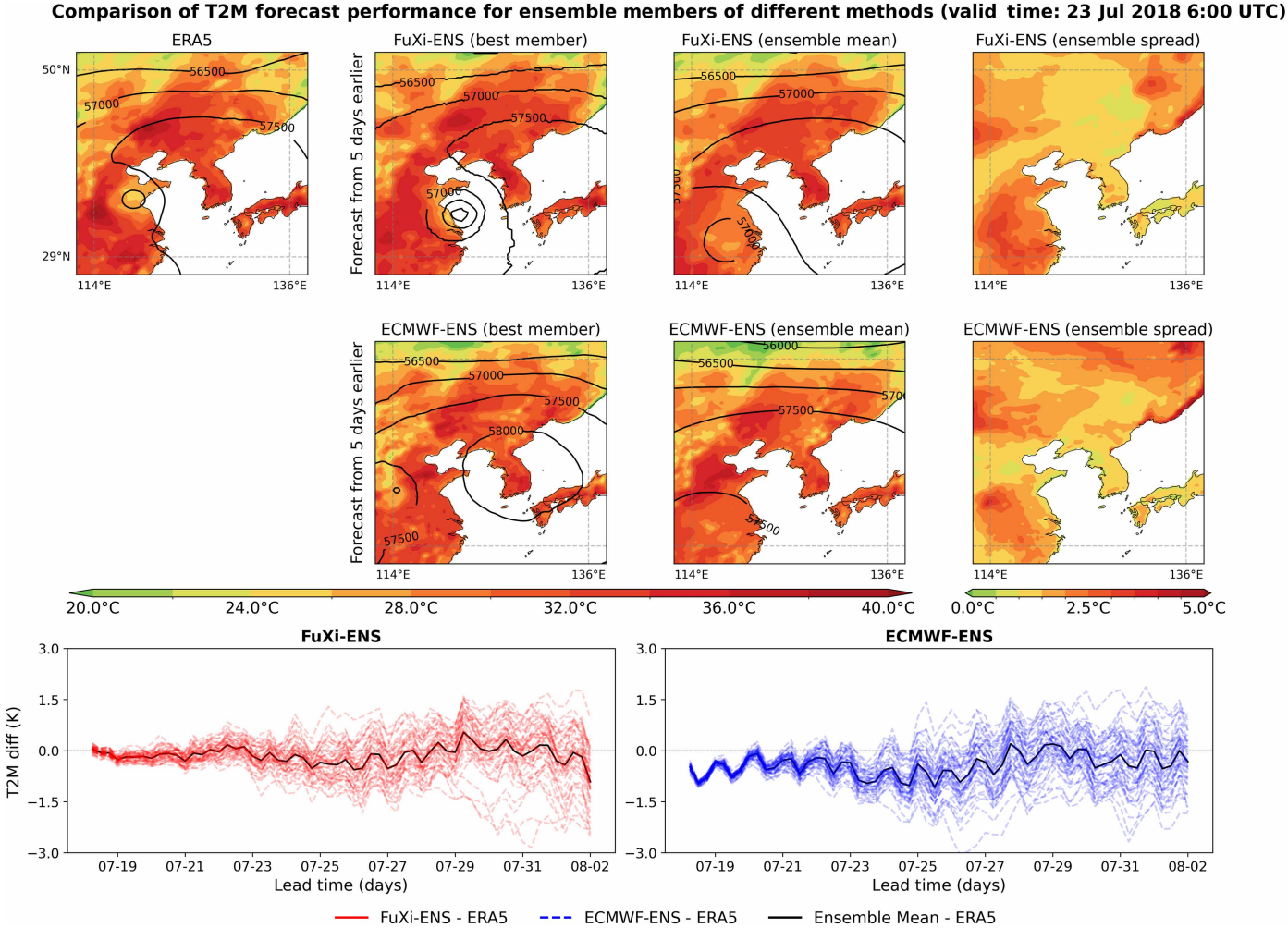


**Fig. 5. Comparison of T2M forecasts during the 2018 Northeast Asia heatwave.** The first and second rows show the spatial distributions of T2M generated by FuXi-ENS and ECMWF-ENS. The first column displays the ERA5 reanalysis data for 6:00 UTC on 23 July 2018. Columns 2 through 4 are predictions from FuXi-ENS (first row) and ECMWF-ENS (second row), showing the best-performing member (second column), ensemble mean (third column), and ensemble spread (fourth column). The top two rows are forecasts made from 5 days earlier. The black contours indicate the 500-hPa geopotential (Z500). The third row shows time series of differences between individual ensemble members and ERA5 (red dashed lines, FuXi-ENS; blue dashed lines, ECMWF-ENS) and between the ensemble mean and ERA5 (black solid lines).

member and ensemble mean of both FuXi-ENS and ECMWF-ENS successfully capture the two centers of regional maximum temperature, located to the north and southwest of Bohai Bay. However, FuXi-ENS demonstrates a lower RMSE over the land area compared to ECMWF-ENS (2.36 K for FuXi-ENS versus 2.76 K for ECMWF-ENS in terms of ensemble mean RMSE). This improvement may be linked to the enhanced accuracy of both the best member and ensemble mean of FuXi-ENS in predicting the strength and pattern of the subtropical high over the Northeast Asia, as well as the anomalous cyclone in the yellow sea, 5 days in advance compared to ECMWF-ENS.

For the worst-performing members (not shown), ECMWF-ENS exhibits larger discrepancies in the position and strength of regional maximum T2M relative to its best member, particularly in northeastern China. This may contribute to a slightly larger ensemble spread for ECMWF-ENS in this region compared to FuXi-ENS. Nonetheless, FuXi-ENS presents a larger ensemble spread in the high-temperature region southwest of Bohai Bay. Overall, both EPSs exhibit similar patterns of T2M spread, with relatively lower spread in regions of high temperature and higher spread elsewhere. In addition to spatial variability, both FuXi-ENS and ECMWF-ENS demonstrate comparable temporal spread among members for spatial-mean T2M forecasts. The daily variation in T2M forecast errors relative to observations also shows consistency between FuXi-ENS and ECMWF-ENS. Additionally, fig. S18 illustrates that FuXi-ENS performs better in predicting the start and end of this heatwave event.

Beyond traditional ensemble verification metrics, it is essential to evaluate how ensemble forecast systems capture underlying physical relationships in space and time. We use ensemble sensitivity analysis (*48*, *49*), using time-lagged ensemble correlations to examine whether FuXi-ENS effectively represents the dynamical processes driving the heatwave development. Given the demonstrated efficacy of ECMWF-ENS in identifying dynamic regimes (*50*), we focus on a comparative analysis between FuXi-ENS and ECMWF-ENS.

The time-lagged ensemble correlation between 246-hour T2M forecasts over a reference region (116°E to 118°E longitude and 39°S to 41°S latitude; marked by a black dot) and Z500 forecasts from 4 and 1 days prior, all from forecasts initialized at 00:00 UTC on 1 July 2018, is presented in fig. S19. Positive correlations indicate that an increase (or decrease) in Z500 in specific regions leads to an increase (or decrease) in T2M at the target location. At both 150-hour and 222-hour lead times, FuXi-ENS and ECMWF-ENS exhibit similar spatial correlation patterns. This demonstrates that the ML-based FuXi-ENS effectively reproduces the process-oriented spatiotemporal covariance captured by the physics-based ECMWF-ENS. Specifically, during the 6- to 10-day forecasts window, rising T2M near the black dot is associated with a dipole pattern in Z500 at day 6 (150 hours): negative correlations to the west and positive correlations to the east. This suggests a shift in the dominant flow pattern from northwesterly to southwesterly, enhancing horizontal warm-air advection and contributing to the warming trend.

By 222 hours, both ensembles exhibit a reversal in the correlation dipole (positive to the west and negative to the east), indicating a continuation of the warm advection regime and further temperature increases at the target location. This process-oriented assessment of ensemble covariance demonstrates that FuXi-ENS performs comparably to ECMWF-ENS in capturing synoptic-scale dynamical processes relevant to the heatwave evolution.

## DISCUSSION

In recent years, ML models have shown great potential in weather forecasting, often outperforming leading NWP models in deterministic forecasts. However, challenges remain in ensemble forecasting, which is crucial for estimating the likelihood of future states, especially for extreme weather events. As lead time increases, forecast accuracy decreases due to the chaotic nature of weather systems, making uncertainty quantification critical. Ensemble forecasts address this by analyzing the spread of predictions among ensemble members, a feature that has become increasingly important as climate change intensifies extreme weather events like TCs.

Recent ML models, such as Google's SEEDS, incorporate additional input data like NWP ensemble members, complicating their operational implementation. Despite superior performance in some metrics compared to ECMWF-ENS, SEEDS is limited by its coarse 2° spatial resolution, and both SEEDS and GenCast, which are based on diffusion models, require more computational steps, leading to slower processing times compared to other ML models. This study introduces FuXi-ENS, an ML-based medium-range ensemble weather forecasting model with a probability-oriented loss function. A key innovation of FuXi-ENS is the integration of the CRPS, for optimizing probability distribution, with KL divergence in its loss function, which improves forecast accuracy over classical VAE models using *L1* and KL loss. By adding flow-dependent perturbations derived from this innovative loss function both at the initial conditions and at each forecast step, further improve its performance. FuXi-ENS can produce 6-hourly forecasts up to 15 days for 5 key upper-air atmospheric variables at 13 pressure levels, as well as 13 surface variables. Comprehensive evaluation demonstrates that FuXi-ENS outperforms ECMWF-ENS in typical forecast metrics, such as RMSE, ACC, CRPS, ROCASS, and BS scores across various variables. Case studies, such as probabilistic TC track forecasts and the 2018 Northeast Asia heatwave, further demonstrate its effectiveness. Compared to conventional NWP ensembles and diffusion-based ML models, FuXi-ENS is both faster and more resource-efficient, producing a 15-day forecast in ~10 s per ensemble member on an Nvidia A100 GPU. In addition to comparisons with ECMWF-ENS, we also compare the performance of FuXi-ENS with the state-of-the-art diffusion-based ensemble model, GenCast (see section S10). This comparison focuses on two key aspects: overall forecast skill and the validity of physical relationships. Both models show overall comparable forecast skill in ensemble mean and probabilistic forecasts, with variations depending on the specific variable. Notably, both FuXi-ENS and GenCast effectively capture fundamental atmospheric physical relationships, particularly under typical weather regimes such as mid-latitude westerly troughs and synoptic-scale frontal zone.

Despite these promising results, FuXi-ENS has certain limitations compared to ECMWF-ENS that need to be addressed in future development. One key issue is the underestimation of ensemble spread in FuXi-ENS relative to ECMWF-ENS and GenCast. ECMWF-ENS uses optimally growing perturbation methods for ensemble initialization, which have not yet been considered for FuXi-ENS. Furthermore, the development of stochastic perturbations in ML-based models is still in its early stages. While FuXi-ENS has made initial efforts to introduce such perturbations, further refinement is necessary. Continued investigation into these aspects will be essential for improving ML-based ensemble forecasting, particularly in capturing extreme events and accurately representing forecast probability distributions. Future enhancements to FuXi-ENS could include

end-to-end training of the forecasting and perturbation models, replacing the univariate with a multivariate Gaussian perturbation distribution, and incorporating model parameter perturbations, such as in Google's functional generative network (FGN) (*51*). The objective is to increase ensemble spread through improved design while simultaneously reducing overall forecast error. Calibration offers another promising avenue, applicable not only to FuXi-ENS but also to other ML-based systems (e.g., GenCast) and traditional numerical models, with the potential to improve reliability and further strengthen forecasting performance.

As ML continues to play an increasingly important role in weather forecasting, particularly for extreme weather events where uncertainty quantification is crucial, FuXi-ENS emerges as a powerful and efficient tool. Moreover, by overcoming the computational constraints of traditional numerical methods, FuXi-ENS could enable the generation of thousands of ensemble members. This could offer considerable potential for ensemble-based data assimilation by providing more accurate and flow-dependent estimates of background error covariances (*52*, *53*).

## MATERIALS AND METHODS

### Data

ERA5, the fifth iteration of the ECMWF reanalysis dataset, provides a comprehensive archive of surface and upper-air variables with a temporal resolution of 1 hour and a horizontal resolution of ~31 km, covering data from January 1950 to the present (*54*). Known for its accuracy and extensive coverage, ERA5 is regarded as the most reliable reanalysis dataset available. For this study, we used the 6-hourly ERA5 reanalysis dataset a spatial resolution of 0.25° (721 × 1440 latitude-longitude grid points).

The FuXi-ENS model forecasts 78 variables, including 5 upper-air atmospheric variables across 13 pressure levels (50, 100, 150, 200, 250, 300, 400, 500, 600, 700, 850, 925, and 1000 hPa) and 13 surface variables. The upper-air variables are geopotential (Z), temperature (T), *u* component of wind (U), *v* component of wind (V), and specific humidity (Q). Surface variables include T2M, 2-m dew-point temperature (D2M), sea surface temperature (SST), 10-m *u* wind component (U10M), 10-m *v* wind component (V10M), 100-m *u* wind component (U100M), 100-m *v* wind component (V100M), mean sea-level pressure (MSL), surface net solar radiation (SNSR), surface net solar radiation downward (SSRD), total sky direct solar radiation at surface (FDIR), top net thermal radiation (TTR), and total precipitation (TP). Table 1 lists these variables and their abbreviations. Variables such as U100M, V100M, SSR, SSRD, and FDIR are particularly relevant for wind and solar energy forecasting.

FuXi-ENS was trained on 15 years of data (2002–2016), with 1 year (2017) for validation and 1 year (2018) for testing. Additional evaluations for TP predictions are provided in the supplementary material.

The ECMWF EPS, operational since 1992, is the leader in medium-range and subseasonal-to-seasonal ensemble forecasting. ECMWF-ENS consists of 51 members: one control forecast generated from the optimal estimate of initial conditions and 50 perturbed forecasts, produced by varying initial conditions and model physics. As of 27 June 2023, its spatial resolution was upgraded to 9 km. For this study, we used ECMWF-ENS forecasts from 2018 at a 0.25° resolution. For validation, the CRPS and BS were calculated using all 51 ensemble members. The "Results" section focuses on the analysis of 500-hPa geopotential (Z500), 850-hPa wind speed (WS850), and T2M. Additional variables, including 850-hPa temperature (T850), MSL, and 10-m wind speed (WS10M), are discussed in the Supplementary Materials.

We assessed TC forecasts using the IBTrACS (*55*, *56*) dataset from the National Oceanic and Atmospheric Administration (NOAA) as the reference. IBTrACS provides global TC tracks at 6-hourly intervals, recording each TC's position with latitude and longitude coordinates. We applied a modified ECMWF TC tracker algorithm (see section S2) (*57*) to the ECMWF and FuXi-ENS ensemble members, as well as to the ERA5 dataset, for TC track extraction and evaluation. To ensure statistical robustness, we excluded forecasts where fewer than two-thirds of the ensemble members detected TCs from our probabilistic TC track analysis (*45*).

### Data preprocessing

*z*-score normalization is used to normalize all input and output variables, with the exception of TP. For upper-air variables, the means and SD are computed separately at each pressure levels using only the training dataset. Before model training, TP is transformed by taking the natural logarithm to enhance the uniformity of its mean and variance across the training data. The transformation is defined as

$$\mathrm{TP}' = \ln(\mathrm{TP} + c) \tag{1}$$

where TP and TP′ represent the original and normalized values, respectively. The constant *c*, introduced to address zero values (*58*), is set to $c = 1$ in this study.

The SST dataset contains not a number (NaN) values over land areas. During preprocessing, all grid points with NaNs across the 15-year training period (2002–2016) were identified and excluded from the loss calculations during model training.

### FuXi-ENS model

Previous ML models for medium-range weather forecasting typically use deterministic encoder-decoder architectures (*22*–*24*, *59*), focusing on minimizing the RMSE against ERA5 reanalysis data. However, these models often fail to account for forecast uncertainty, which is particularly critical for predicting extreme weather events. Ensemble forecasting presents an additional challenge for ML models, as it lacks a definitive "ground truth" for introducing perturbations, which are essential for estimating uncertainties.

Inspired by conventional NWP models that introduce perturbations to both initial conditions and forecast steps to account for errors and uncertainties, we developed the FuXi-ENS model, an autoregressive ML model specifically designed for ensemble weather forecasting. The model incorporates a VAE (*60*–*62*), a probabilistic generative model that encodes input data into a Gaussian distribution, from which perturbations are sampled. This sampling process introduces randomness and variability, providing a direct measure of uncertainty, with the spread of the distribution reflects the model's confidence in the latent representation of the data.

The FuXi-ENS model has two primary components: a perturbation model and a forecasting model (see Fig. 6). The perturbation model, based on a VAE, transforms input data into a Gaussian distribution that captures the probabilistic characteristics of the data. The forecasting model, using an encoder-decoder framework, generates ensemble forecasts by sampling multiple times from this Gaussian distribution, with each sample representing one ensemble member. This process is repeated at each forecasting step to improve

**Table 1. A summary of all the input and output variables.** The "Type" indicates whether the variable is a time-varying variable, including upper-air, surface, and geographical variables, or a temporal variable. The "Full name" and "Abbreviation" columns refer to the complete name of each variable and their corresponding abbreviations in this paper. The "Role" column clarifies whether each variable serves as both an input and an output or is solely used as an input by our model.

| Type | Full name | Abbreviation | Role |
|---|---|---|---|
| Upper-air variables | Geopotential | Z | Input and output |
| | Temperature | T | Input and output |
| | *u* component of wind | U | Input and output |
| | *v* component of wind | V | Input and output |
| | Specific humidity | Q | Input and output |
| Surface variables | 2-m temperature | T2M | Input and output |
| | 2-m dew-point temperature | D2M | Input and output |
| | Sea surface temperature | SST | Input and output |
| | 10-m *u* wind component | U10M | Input and output |
| | 10-m *v* wind component | V10M | Input and output |
| | 100-m *u* wind component | U100M | Input and output |
| | 100-m *v* wind component | V100M | Input and output |
| | Mean sea-level pressure | MSL | Input and output |
| | Surface net solar radiation | SNSR | Input and output |
| | Surface net solar radiation Downward | SSRD | Input and output |
| | Total sky direct solar radiation at surface | FDIR | Input and output |
| | Top net thermal radiation | TTR | Input and output |
| | Total precipitation | TP | Input and output |
| Geographical | Orography | OR | Input |
| | Land-sea mask | LSM | Input |
| | Latitude | LAT | Input |
| | Longitude | LON | Input |
| Temporal | Hour of day | HOUR | Input |
| | Day of year | DOY | Input |
| | Step | STEP | Input |

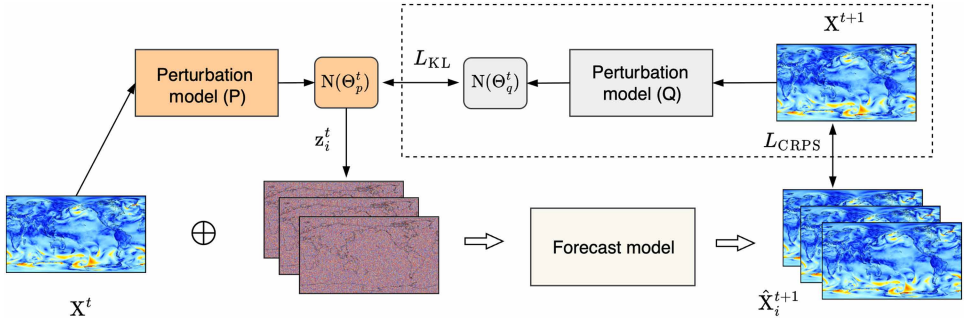


**Fig. 6. Schematic diagram of the structures of the FuXi-ENS model.**

uncertainty estimation and enhance the model's effectiveness in ensemble forecasting. To clarify the methodology, this approach is analogous to traditional ensemble forecasting techniques: the forecasting model generates deterministic forecasts from initial conditions, while the perturbation model introduces flow-dependent variations to account for uncertainties in both initial conditions and predictions at each forecast step.

The input to FuXi-ENS is a data cube with dimensions 2 by 78 by 721 by 1440, representing meteorological variables from two previous time steps ($t-1$ and $t$), the number of variables ($C$), and grid points along latitude ($H$) and longitude ($W$), respectively. This data cube is initially reshaped to ($2C$) by $H$ by $W$ (156 by 721 by 1440). To compress the input data into a smaller, more abstract representation, the perturbation model first reduces the dimensions of this meteorological data

cube to 2*C* by 90 by 180 using a two-dimensional (2D) convolution layer with a kernel size of 8 and a stride of 8. Simultaneously, the geographical and temporal variables are processed by an identical 2D convolution layer. The resulting output is then concatenated with the meteorological data and passes through 14 Swin Transformer (*63*) blocks, each using an 18 by 18 window. Subsequently, the data are restored to its original dimensions via a 2D transposed convolution layer, also with a kernel size and stride of 8 (*64*). The final output is a Gaussian distribution $[\mathrm{N}(\Theta^t{}_p)]$, defined by a mean matrix $\mu^t$ and a covariance matrix $\sigma^t$, both of size 156 by 721 by 1440. A perturbation vector $\mathbf{z}^t$ is sampled from $\mathrm{N}(\Theta^t{}_p)$ and added to the input data, yielding perturbed initial conditions ($\tilde{\mathbf{X}}^t = \mathbf{X}^t + \mathbf{z}^t$). These perturbed initial conditions are fed into the forecasting model, which applies an 8 by 8 2D convolution layer, followed by 40 transformer blocks and an 8 by 8 2D transposed convolution layer, to generate the final ensemble forecasts $\hat{\mathbf{X}}^{t+1}$. The ensemble size is determined by the number of samples drawn from the Gaussian distribution $\mathrm{N}(\Theta^t{}_p)$.

### FuXi-ENS model training

The FuXi-ENS model was developed using the Pytorch framework (*65*). Its training consists of two stages: (i) pretraining the forecasting model and (ii) jointly training the forecasting and perturbation models. In the first stage, the forecasting model is pretrained to predict a single 6-hour step, following a procedure similar to that used in FuXi (*24*). A latitude-weighted *L1* loss is used, defined as

$$L1 = \frac{1}{C \times H \times W} \sum_{c=1}^{C} \sum_{i=1}^{H} \sum_{j=1}^{W} a_i \left| \hat{\mathbf{X}}^{t+1}_{c,i,j} - \mathbf{X}^{t+1}_{c,i,j} \right| \tag{2}$$

where $a_i = H \times \cos\Phi_i / \sum_{i=1}^{H} \cos\Phi_i$ is a latitude-dependent weight at latitude $\Phi_i$, which decreases with increasing latitude. The forecasting model is trained for 60,000 iterations on eight Nvidia A100 GPUs, with a batch size of 1 per GPU. The AdamW (*66*, *67*) optimizer is used with $\beta_1 = 0.9$ and $\beta_2 = 0.95$, an initial learning rate of $2.5 \times 10^{-4}$, and a weight decay coefficient of 0.1.

In the second stage, the forecasting and perturbation models are trained jointly to minimize a combination of the CRPS and the KL divergence loss. CRPS is a key metric for assessing ensemble forecast quality, while the KL loss serves as a regularization term that aligns the predicted Gaussian distribution $[\mathrm{N}(\Theta^t{}_p)]$ with the target distribution, thus improving the uncertainty representation. A key challenge is reconciling discrepancies between these distributions due to prediction errors, which is addressed through a knowledge distillation strategy. A perturbation model (Q) transforms target data into a Gaussian distribution to supervise the perturbation model P. This supervision minimizes the KL loss ($L_{\mathrm{KL}}$), which quantifies the difference between the distributions of Q and P. During training, perturbation model Q processes target data from two previous time steps, $\mathbf{X}^t$ and $\mathbf{X}^{t+1}$, to generate a Gaussian distribution $[\mathrm{N}(\Theta^t{}_q)]$ similar to that of model P. Intermediate perturbation vectors are sampled from model Q's distribution during training, and from model P's distribution $[\mathrm{N}(\Theta^t{}_p)]$ during testing. Meanwhile, the CRPS loss is calculated between the ensemble forecast ($\hat{\mathbf{X}}^{t+1}$) and the target data $\mathbf{X}^{t+1}$. The overall loss function balances the CRPS loss and KL loss terms

$$L = L_{\mathrm{CRPS}}\left(\hat{\mathbf{X}}^{t+1}_i, \mathbf{X}^{t+1}_i\right) + \lambda L_{\mathrm{KL}}\left[\mathrm{N}(\Theta^t{}_p), \mathrm{N}(\Theta^t{}_q)\right] \tag{3}$$

where λ is a tunable coefficient, set to $1 \times 10^{-4}$, balancing $L_{\mathrm{KL}}$ and $L_{\mathrm{CRPS}}$ loss terms. This design ensures that perturbation vectors closely approximate the true data distribution and improve the CRPS of the ensemble forecasts.

Joint training follows an autoregressive training regime with a curriculum training schedule (*23*), in which the number of autoregressive steps increases from 1 to 3, with each step consisting of 3,000 training iterations. The learning rate is fixed at $4.5 \times 10^{-6}$, and the batch size is 1 per GPU. Training is conducted on a cluster of 48 Nvidia A100 GPUs, with each GPU generating a single ensemble member, thereby producing a 48-member ensemble. Both the FuXi-ENS model and the initial conditions are replicated across all 48 GPUs, facilitating parallel generation of ensemble forecasts.

For testing, a single GPU can generate one member at a time, which can be repeated 48 times to construct a 48-member ensemble. At each forecast step, the perturbation model P outputs a Gaussian distribution, from which the perturbation vector $\mathbf{z}^t$ is randomly sampled and applied to perturb the input. This perturbed input is then used by the forecasting model to produce a 6-hour forecast, which is fed back into both models for the next time step. This process is repeated autoregressively to generate 15-day forecasts at a 6-hour resolution. Alternatively, inference can be further accelerated by parallelizing across multiple GPUs, as each ensemble member is generated independently.

As illustrated in fig. S24, the FuXi-ENS model, trained with CRPS, outperforms the model trained with *L1* loss not only in terms of CRPS but also in RMSE, ensemble spread, and SSR. The loss function for FuXi-ENS, trained with a combination of *L1* and KL loss, is defined as follows

$$L = L1\left(\hat{\mathbf{X}}^{t+1}, \mathbf{X}^{t+1}\right) + \lambda L_{\mathrm{KL}}\left(P^t, Q^t\right) \tag{4}$$

where *L1* replaces the CRPS loss.

## Supplementary Materials

**This PDF file includes:**

Supplementary Text
Figs. S1 to S31
Table S1

## REFERENCES AND NOTES


1. E. N. Lorenz, Deterministic nonperiodic flow. *J. Atmos. Sci.* **200**, 130–148 (1963).
2. National Research Council, Division on Earth and Life Studies, Board on Atmospheric Sciences and Climate, Committee on Estimating and Communicating Uncertainty in Weather and Climate Forecasts, *Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts* (National Academies Press, 2006).
3. S. Scher, G. Messori, Predicting weather forecast uncertainty with machine learning. *Q. J. R. Meteorol. Soc.* **1440**, 2830–2841 (2018).
4. C. Calvo-Olivera, Á. M. Guerrero-Higueras, J. Lorenzana, E. García-Ortega, Real-time evaluation of the uncertainty in weather forecasts through machine learning-based models. *Water Resour. Manage* **380**, 2455–2470 (2024).
5. T. N. Palmer, The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Q. J. R. Meteorol. Soc.* **1280**, 747–774 (2002).
6. L. Goodarzi, M. E. Banihabib, A. Roozbahani, A decision-making model for flood warning system based on ensemble forecasts. *J. Hydrol.* **573**, 207–219 (2019).
7. S. Sperati, S. Alessandrini, L. D. Monache, An application of the ECMWF Ensemble Prediction System for short-term solar power forecasting. *Sol. Energy* **133**, 437–450 (2016).
8. Y.-K. Wu, P.-E. Su, T.-Y. Wu, J.-S. Hong, M. Y. Hassan, Probabilistic wind-power forecasting using weather ensemble models. *IEEE Trans. Ind. Appl.* **540**, 5609–5620 (2018).

9. B. Zhang, L. Tang, M. Roemer, Probabilistic planning and risk evaluation based on ensemble weather forecasting. *IEEE Trans Autom Sci Eng* **150**, 556–566 (2017).
10. T. Gneiting, A. E. Raftery, Weather forecasting with ensemble methods. *Science* **310**, 248–249 (2005).
11. M. Leutbecher, T. N. Palmer, Ensemble forecasting. *J. Comput. Phys.* **227**, 3515–3539 (2008).
12. F. Molteni, R. Buizza, T. N. Palmer, T. Petroliagis, The ECMWF ensemble prediction system: Methodology and validation. *Q. J. R. Meteorol. Soc.* **122**, 73–119 (1996).
13. R. Buizza, T. N. Palmer, The singular-vector structure of the atmospheric global circulation. *J. Atmos. Sci.* **520**, 1434–1456 (1995).
14. J. Barkmeijer, R. Buizza, T. N. Palmer, 3D-Var Hessian singular vectors and their potential use in the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.* **125**, 2333–2351 (1999).
15. R. Buizza, M. Milleer, T. N. Palmer, Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.* **125**, 2887–2908 (1999).
16. T. N. Palmer, R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G. J. Shutts, M. Steinheimer, A. Weisheimer, "Stochastic parametrization and model uncertainty" (ECMWF Technical Memoranda, European Centre for Medium-Range Weather Forecasts, 2009).
17. R. Buizza, Introduction to the special issue on "25 years of ensemble forecasting". *Q. J. R. Meteorol. Soc.* **145**, 1–11 (2019).
18. M. Leutbecher, Ensemble size: How suboptimal is less than infinity? *Q. J. R. Meteorol. Soc.* **145**, 107–128 (2019).
19. S. T. K. Lang, A. Dawson, M. Diamantakis, P. Dueben, S. Hatfield, M. Leutbecher, T. Palmer, F. Prates, C. D. Roberts, I. Sandu, N. Wedi, More accuracy with less precision. *Q. J. R. Meteorol. Soc.* **147**, 4358–4370 (2021).
20. X. Zhou, Y. Zhu, D. Hou, B. Fu, W. Li, H. Guan, E. Sinsky, W. Kolczynski, X. Xue, Y. Luo, J. Peng, B. Yang, V. Tallapragada, P. Pegion, The development of the NCEP Global Ensemble Forecast System version 12. *Weather Forecast.* **370**, 1069–1084 (2022).
21. J. Pathak, S. Subramanian, P. Harrington, S. Raja, A. Chattopadhyay, M. Mardani, T. Kurth, D. Hall, Z. Li, K. Azizzadenesheli, P. Hassanzadeh, K. Kashinath, A. Anandkumar, FourCastNet: A global data-driven high-resolution weather model using adaptive fourier neural operators. arXiv:2202.11214 (2022).
22. K. Bi, L. Xie, H. Zhang, X. Chen, G. Xiaotao, Q. Tian, Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **619**, 533–538 (2023).
23. R. Lam, A. Sanchez-Gonzalez, M. Willson, P. Wirnsberger, M. Fortunato, F. Alet, S. Ravuri, T. Ewalds, Z. Eaton-Rosen, H. Weihua, A. Merose, S. Hoyer, G. Holland, O. Vinyals, J. Stott, A. Pritzel, S. Mohamed, P. Battaglia, Learning skillful medium-range global weather forecasting. *Science* **382**, 1416–1421 (2023).
24. L. Chen, X. Zhong, F. Zhang, Y. Cheng, X. Yinghui, Y. Qi, H. Li, FuXi: A cascade machine learning forecasting system for 15-day global weather forecast. *npj Clim. Atmos. Sci.* **60**, 190 (2023).
25. S. Lang, M. Alexe, M. Chantry, J. Dramsch, F. Pinault, B. Raoult, M. C. A. Clare, C. Lessig, M. Maier-Gerber, L. Magnusson, Z. B. Bouallègue, A. P. Nemesio, P. D. Dueben, A. Brown, F. Pappenberger, F. Rabier, AIFS - ECMWF's data-driven forecasting system. arXiv:2406.01465 (2024).
26. I. Price, A. Sanchez-Gonzalez, F. Alet, T. R. Andersson, A. El-Kadi, D. Masters, T. Ewalds, J. Stott, S. Mohamed, P. Battaglia, R. Lam, M. Willson, Probabilistic weather forecasting with machine learning. *Nature* **637**, 84–90 (2025).
27. J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, S. Ganguli, Deep unsupervised learning using nonequilibrium thermodynamics. arXiv:1503.03585 (2015).
28. T. Karras, M. Aittala, T. Aila, S. Laine, Elucidating the design space of diffusion-based generative models. *Adv. Neural Inf. Process. Syst.* **35**, 26565–26577 (2022).
29. L. Li, R. Carver, I. Lopez-Gomez, F. Sha, J. Anderson, Generative emulation of weather forecast ensembles with diffusion models. *Sci. Adv.* **10**, eadk4489 (2024).
30. N. D. Brenowitz, Y. Cohen, J. Pathak, A. Mahesh, B. Bonev, T. Kurth, D. R. Durran, P. Harrington, M. S. Pritchard, A practical probabilistic benchmark for AI weather models. arXiv:2401.15305 (2024).
31. H. Yuan, L. Chen, Z. Wang, H. Li, SwinVRNN: A data-driven ensemble forecasting model via learned distribution perturbation. *J. Adv. Model. Earth Syst.* **15**, e2022MS003211 (2023).
32. M. E. Mousavi, J. L. Irish, A. E. Frey, F. Olivera, B. L. Edge, Global warming and hurricanes: The potential impact of hurricane intensification and sea level rise on coastal flooding. *Clim. Chang.* **104**, 575–597 (2011).
33. S. T. K. Lang, M. Leutbecher, S. C. Jones, Impact of perturbation methods in the ECMWF ensemble prediction system on tropical cyclone forecasts. *Quart. J. Roy. Meteorol. Soc.* **138**, 2030–2046 (2012).
34. J. D. Brown, J. Demargne, D.-J. Seo, Y. Liu, The Ensemble Verification System (EVS): A software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. *Environ. Model. Software* **25**, 854–872 (2010).
35. V. Fortin, M. Abaza, F. Anctil, R. Turcotte, Why should ensemble spread match the RMSE of the ensemble mean? *J. Hydrometeorol.* **15**, 1708–1713 (2014).
36. A. H. Murphy, E. S. Epstein, Skill scores and correlation coefficients in model verification. *Mon. Wea. Rev.* **117**, 572–582 (1989).
37. T. M. Hamill, G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau, Y. Zhu, W. Lapenta, NOAA's second-generation global medium-range ensemble reforecast dataset. *Bull. Amer. Meteor. Soc.* **94**, 1553–1565 (2013).
38. S. J. Mason, N. E. Graham, Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance. *Quart. J. Roy. Meteorol. Soc.* **128**, 2145–2166 (2002).
39. H. Hersbach, Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecast.* **15**, 559–570 (2000).
40. G. W. Brier, Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.* **78**, 1–3 (1950).
41. H.-C. Tsai, R. L. Elsberry, Detection of tropical cyclone track changes from the ECMWF ensemble prediction system. *Geophys. Res. Lett.* **40**, 797–801 (2013).
42. C. W. Landsea, J. P. Cangialosi, Have we reached the limits of predictability for tropical cyclone track forecasting? *Bull. Amer. Meteor. Soc.* **99**, 2237–2243 (2018).
43. X. Zhong, L. Chen, J. Liu, C. Lin, Y. Qi, H. Li, FuXi-Extreme: Improving extreme rainfall and wind forecasts with diffusion model. *Sci. China Earth Sci.* **67**, 3696–3708 (2024a).
44. M. DeMaria, J. A. Knaff, M. J. Brennan, D. Brown, R. D. Knabb, R. T. DeMaria, A. Schumacher, C. A. Lauer, D. P. Roberts, C. R. Sampson, P. Santos, D. Sharp, K. A. Winters, Improvements to the operational tropical cyclone wind speed probability model. *Wea. Forecast.* **28**, 586–602 (2013).
45. X. Zhang, G. Chen, Y. Hui, Z. Zeng, Verification of ensemble track forecasts of tropical cyclones during 2014. *Trop. Cyclone Res. Rev.* **4**, 79–87 (2015).
46. World Meteorological Organization, July sees extreme weather with high impacts (2018). https://wmo.int/media/july-sees-extreme-weather-high-impacts.
47. P.-C. Hsu, Y. Yitian Qian, H. M. Liu, Y. Gao, Role of abnormally enhanced MJO over the Western Pacific in the formation and subseasonal predictability of the record-breaking Northeast Asian heatwave in the summer of 2018. *J. Clim.* **33**, 3333–3349 (2020).
48. R. D. Torn, G. J. Hakim, Ensemble-based sensitivity analysis. *Mon. Wea. Rev.* **136**, 663–677 (2008).
49. J. S. Whitaker, G. P. Compo, J.-N. Thépaut, A comparison of variational and ensemble-based data assimilation systems for reanalysis of sparse observations. *Mon. Wea. Rev.* **137**, 1991–1999 (2009).
50. L. Liu, J. Feng, L. Ma, Y. Yang, W. Xiaotian, C. Wang, Ensemble-based sensitivity analysis of track forecasts of typhoon In-fa (2021) without and with model errors in the ECMWF, NCEP, and CMA ensemble prediction systems. *Atmos. Res.* **309**, 1–14 (2024).
51. F. Alet, I. Price, A. El-Kadi, D. Masters, S. Markou, T. R. Andersson, J. Stott, R. Lam, M. Willson, A. Sanchez-Gonzalez, P. Battaglia, Skillful joint probabilistic weather forecasting from marginals. arXiv:2506.10772 (2025).
52. M. Bonavita, E. Hólm, L. Isaksen, M. Fisher, The evolution of the ECMWF hybrid data assimilation system. *Quart. J. Roy. Meteorol. Soc.* **142**, 287–303 (2016).
53. P. L. Houtekamer, F. Zhang, Review of the ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.* **144**, 4489–4532 (2016).
54. H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellan, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, G. De Chiara, P. Dahlgren, D. Dee, M. Diamantakis, R. Dragani, J. Flemming, R. Forbes, M. Fuentes, A. Geer, L. Haimberger, S. Healy, R. J. Hogan, E. Hólm, M. Janisková, S. Keeley, P. Laloyaux, P. Lopez, C. Lupu, G. Radnoti, P. de Rosnay, I. Rozum, F. Vamborg, S. Villaume, J.-N. Thépaut, The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **146**, 1999–2049 (2020).
55. K. R. Knapp, M. C. Kruk, D. H. Levinson, H. J. Diamond, C. J. Neumann, The International Best Track Archive for Climate Stewardship (IBTrACS) unifying tropical cyclone data. *Bull. Amer. Meteor. Soc.* **91**, 363–376 (2010).
56. K. R. Knapp, H. J. Diamond, J. P. Kossin, M. C. Kruk, C. J. Schreck, International Best Track Archive for Climate Stewardship (IBTrACS) project, version 4.01, NOAA National Centers for Environmental Information (2018).
57. V. der Grijn, "Tropical cyclone forecasting at ECMWF: New products and validation" (ECMWF Technical Memoranda, European Centre for Medium-Range Weather Forecasts, 2002).
58. G. Lien, E. Kalnay, T. Miyoshi, G. J. Huffman, Statistical properties of global precipitation in the NCEP GFS model and TMPA observations for data assimilation. *Mon. Wea. Rev.* **144**, 663–679 (2016).
59. L. Olivetti, G. Messori, Advances and prospects of deep learning for medium-range extreme weather forecasting. *Geosci. Model Dev.* **17**, 2347–2358 (2024).
60. C. Doersch, Tutorial on variational autoencoders. arXiv:1606.05908 (2016).
61. T. Zhao, R. Zhao, M. Eskenazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. arXiv:1703.10960 (2017).
62. D. P. Kingma, M. Welling, An introduction to variational autoencoders. *Found. Trends Mach. Learn.* **12**, 307–392 (2019).
63. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vission (ICCV)* (IEEE, 2021), pp. 9992–10022.

64. M. D. Zeiler, D. Krishnan, G. W. Taylor, R. Fergus, "Deconvolutional networks," in *2010 IEEE Computer Society Conference on Computer Vission and Pattern Recognition* (IEEE, San Francisco, CA, USA, 2010), pp. 2528–2535.
65. A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, "Automatic differentiation in PyTorch," in *NIPS 2017 Workshop on Autodiff* (NIPS, 2017).
66. D. P. Kingma, J. Ba, Adam: A method for stochastic optimization. arXiv:1412.6980 (2014).
67. I. Loshchilov, F. Hutter, Decoupled weight decay regularization. arXiv:1711.05101 (2017).
68. X. Zhong, L. Chen, H. Li, R. Buizza, J. Liu, J. Feng, Z. Zhu, X. Fan, K. Dai, J.-J. Luo, J. Wu, B. Lu, FuXi-ENS data, version 1.0, Zenodo (2025); https://zenodo.org/uploads/15171712.
69. X. Zhong, L. Chen, H. Li, R. Buizza, J. Liu, J. Feng, Z. Zhu, X. Fan, K. Dai, J.-J. Luo, J. Wu, B. Lu. FuXi-ENS model, version 1.0, Zenodo (2025); https://zenodo.org/records/15124541.

**Acknowledgments:** We extend our sincere appreciation to the researchers at ECMWF and Google for invaluable contributions in collecting, archiving, disseminating, and maintaining the ERA5 reanalysis dataset and ECMWF-ENS. The computations in this research were performed using the CFFF platform of Fudan University. **Funding:** This work was supported by the National Key R&D Program of China under grant 2021YFA0718000, the National Natural Science Foundation of China under grant 42175052, the Postdoctoral Fellowship Program of CPSF under grant number GZB20240154, and the China Meteorological Administration Joint Research Project for Meteorological Capacity Improvement (24NLTSZD03). **Author contributions:** Conceptualization: X.Z., L.C., H.L., K.D., J.-j.L., and B.L. Methodology: X.Z., L.C., H.L., and B.L. Investigation: X.Z., L.C., R.B., J.L., J.F., Z.Z., and B.L. Formal analysis: X.Z., L.C., J.L., J.F., and Z.Z. Software: X.Z., L.C., H.L., and Z.Z. Data curation: X.Z., L.C., J.L., and Z.Z. Visualization: X.Z., R.B., J.L., Z.Z., and X.F. Project administration: X.Z. and H.L. Supervision: H.L. and B.L. Writing—original draft: X.Z., J.F., Z.Z., K.D., and J.-j.L. Writing—review and editing: X.Z., L.C., R.B., J.L., J.F., K.D., J.-j.L., J.W., and B.L. Funding acquisition: X.Z., H.L., and B.L. Validation: X.Z., L.C., R.B., J.F., Z.Z., K.D., and J.W. Resources: L.C., R.B., J.L., Z.Z., K.D., and B.L. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. The ERA5 dataset subset used in this study was downloaded from Copernicus Climate Data (CDS) at https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels?tab=overview for ERA5 data on single levels and https://cds.climate.copernicus.eu/datasets/reanalysis-era5-pressure-levels?tab=overview for ERA5 data on pressure levels. ECMWF-ENS means and SD are available at https://apps.ecmwf.int/archive-catalogue/?type=em&class=od&stream=enfo&expver=1 ECMWF-ENS data were obtained from the WeatherBench 2 public cloud storage at the link https://console.cloud.google.com/storage/browser/weatherbench2/datasets/ifs_ens. FuXi-ENS data are available at https://zenodo.org/records/15171712 (*68*). The GenCast 2019 historical forecasts used in this study are available at https://developers.google.com/earth-engine/datasets/catalog/projects_gcp-public-data-weathernext_assets_126478713_1_0 (*26*). The code and model for running the FuXi-ENS model in this study is available at https://zenodo.org/records/15124541 (*69*). The code and model for running GenCast model in this study is available from https://github.com/ecmwf-lab/ai-models-gencast?tab=readme-ov-file (*26*). The xskillscore Python package, available at https://github.com/xarray-contrib/xskillscore/, is an open-source tool developed independently of this work. None of the authors of this manuscript are affiliated with its development or maintenance.

Submitted 1 November 2024
Accepted 2 October 2025
Published 31 October 2025
10.1126/sciadv.adu2854

Science Advances

# FuXi-ENS: A machine learning model for efficient and accurate ensemble weather prediction

Xiaohui Zhong, Lei Chen, Hao Li, Roberto Buizza, Jun Liu, Jie Feng, Zijian Zhu, Xu Fan, Kan Dai, Jing-jia Luo, Jie Wu, and Bo Lu

*Sci. Adv.* **11** (44), eadu2854. DOI: 10.1126/sciadv.adu2854