



A fast physics-based perturbation generator of machine learning weather model for efficient ensemble forecasts of tropical cyclone track



Jingchen Pu¹, Mu Mu^{1,2,3}, Jie Feng^{1,2,4} ✉, Xiaohui Zhong⁵ & Hao Li⁵

Traditional ensemble forecasting based on numerical weather prediction (NWP) models, is constrained by the need for massive computational resources, resulting in limited ensemble sizes. Although emerging artificial intelligence (AI)-based weather models offer high forecast accuracy and improved computational efficiency, they still face considerable challenges in ensemble forecasting applications, due to the unclear error growth dynamic and the lack of suitable ensemble methods in AI-based models. In this study, we propose a fast, physics-constrained perturbation scheme through the self-evolution dynamics of an AI-based weather model for ensemble forecasting of tropical cyclones (TCs). These initial perturbations are conditioned on specific amplitude and spatial characteristics, exhibiting physically reasonable dynamical growth and spatial covariance. Based on this perturbation scheme, the TC track ensemble forecasts within the AI-based model significantly outperform those from the European Centre for Medium-Range Weather Forecasts (ECMWF) for both deterministic and probabilistic metrics. Notably, we conduct TC track forecasts with 2000 members for the first time, achieving further enhanced forecast skills in probability distribution and extreme scenarios of TC movement.

Tropical cyclones (TCs) are among the most destructive weather phenomena impacting human lives and property in tropical regions. TCs can cause multiple disasters, such as wind gusts, heavy rainfall, and storm surges. Therefore, enhancing predictive capabilities for TCs, especially their tracks, is crucial for mitigating related damages^{1,2}. Previous research indicates that TC movement is influenced by complex, multi-scale factors, including large-scale synoptic systems^{3,4}, vortex structure^{5,6}, sea surface temperature^{7,8}, and cloud convection effects⁹. These multi-scale factors introduce significant uncertainties in predicting TC tracks^{10,11}. A deterministic forecast provides only one possible future state of a TC. In contrast, ensemble forecasts, which comprise a set of varied predictions, can offer probabilistic information about future states and help estimate forecast uncertainty¹². Thus, ensemble forecasting has emerged as an effective method for improving the accuracy of TC track predictions^{13–15}.

Major operational forecast centers worldwide have developed their own ensemble prediction systems (EPS) using global or regional numerical weather prediction (NWP) models^{16–19}, which are crucial for providing predictions for TCs. Advances in observing systems, data assimilation, ensemble generation schemes, and NWP model performance have consistently improved the accuracy of these EPSs^{20,21}. However, computational costs impose significant limitations, restricting operational EPSs to a limited number of ensemble members, typically ranging from 10 to 50. Given that this range is significantly lower than the degrees of freedom in NWP models (generally on the order of 10^8), these systems can only sample a subspace of the atmospheric variable phase space. This restriction may lead to sub-optimal performance in representing the probability distribution of variables and in quantifying forecast uncertainties^{22,23}. Despite the urgent operational need to develop higher-resolution NWP models for better

¹Department of Atmospheric and Oceanic Sciences and Institute of Atmospheric Sciences, Fudan University, Shanghai, China. ²Shanghai Key Laboratory of Ocean-land-atmosphere Boundary Dynamics and Climate Change, Shanghai, China. ³Shanghai Frontiers Science Center of Atmosphere–Ocean Interaction, Shanghai, China. ⁴Shanghai Academy of Artificial Intelligence for Science, Shanghai, China. ⁵Artificial Intelligence Innovation and Incubation Institute, Fudan University, Shanghai, China. ✉e-mail: fengjie@fudan.edu.cn

resolving extreme weather events, significantly increasing the number of ensemble members remains a formidable challenge in operational settings.

In recent years, artificial intelligence (AI) technology has advanced rapidly and has been widely applied to meteorological forecasting^{24,25}, providing a swift and effective alternative to traditional NWP²⁶. AI-based weather forecast models are developed by training on the historical long-range observations of three-dimensional atmospheric variables, typically using the fifth-generation ECMWF reanalysis (ERA5) data. This process employs a machine learning (ML) framework, mostly based on transformers. Leveraging graphics processing unit (GPU) acceleration, these data-driven models can produce 15-day global weather forecasts within seconds, achieving speeds over 10,000 times faster than those of traditional NWP²⁷. A significant milestone was reached with the development of the Pangu model²⁷, which was the first to demonstrate overall more accurate medium-range forecasts than the most advanced NWP model, the Integrated Forecast System (IFS) of the ECMWF. Following this, a series of AI weather models with higher accuracy emerged, including GraphCast²⁸, FuXi²⁹, and FengWu³⁰, each utilizing an encoder-decoder framework but with distinct architectures. Notably, some of these data-driven models have been shown to provide more accurate deterministic TC track forecasts on average than the IFS^{27,28}. Given their superior forecasting skill, significantly lower computational cost, and faster processing speed, it is worthwhile to explore the possibility of generating thousands of TC ensemble forecasts and their potential benefits for TC predictions, a feat that is currently unfeasible with traditional NWP frameworks.

The core principle of traditional ensemble forecasting, based on the NWP model, lies in sampling the uncertainties inherent in forecasts, including those caused by initial and model uncertainties^{31,32}. Although major operational EPSs employ distinct schemes for generating these perturbations in initial analyses and numerical models, their underlying principles align: to identify and integrate key sources of uncertainties^{33,34}. These physically constrained perturbations aim to improve the performance of the evolving perturbations in capturing forecast errors, thereby producing more representative ensemble forecast members that more accurately reflect the true states of the atmosphere^{16,17}.

In contrast to well-established traditional EPSs, AI-based ensemble forecasting remains in its early stages. A fundamental challenge in this field is the limited understanding of the dynamic characteristics of data-driven models. A few studies, such as those by Bi et al.²⁷ and Chen et al.²⁹, have introduced random Perlin (Perlin noise is a type of gradient noise developed by Ken Perlin in 1983. It has many applications, including but not limited to: procedurally generating terrain, applying pseudo-random changes to a variable, and assisting in the creation of image textures) noise into the initial conditions of data-driven models for ensemble generation. Although their ensemble mean forecasts exhibit lower root-mean-square errors (RMSE) compared to the control forecasts, they did not thoroughly investigate the spatial characteristics and dynamic evolution of ensemble perturbations. Moreover, the randomly generated Perlin noise lacks flow-dependent properties, which has been identified as detrimental to the performance of ensemble forecasting^{12,35,36}. Other studies have explored the application of classical dynamically generated perturbations to AI models. For example, Brenowitz et al.³⁷ suggested that lag forecasts could serve as a practical benchmark for AI weather models. Scher and Messori³⁸ and Bulte et al.³⁹ tested the singular vector and Gaussian noise in their quantification of AI forecast uncertainties. Although some of these approaches reported competitive results for ensemble mean forecasts compared to operational forecasts, the growth of perturbations was rarely studied, resulting in the empirical design of ensemble methods and flawed results for metrics such as ensemble spread.

Researchers have increasingly recognized the importance of analyzing the perturbation growth and physical consistency in AI models. Selz and Craig⁴⁰ explored the dynamic sensitivity of the Pangu model to initial perturbations. They highlighted that data-driven models might not effectively capture the upscale evolution of small-scale perturbations, most often referred to as the “butterfly effect”. Bonavita⁴¹ analyzed the physical balances

and spectral characteristics of forecasts from data-driven models. He argued that these models may lack physical consistency across variables and struggle to simulate mesoscale systems. These findings collectively imply that adapting traditional perturbation initialization schemes to AI weather models might be challenging.

In addition to traditional approaches that perturb initial conditions for AI ensemble forecasting, recent studies have incorporated advanced ML algorithms into ensemble forecasting. Zhong et al.⁴² constructed a variational autoencoder scheme that transforms forecast data at each iterative step into a Gaussian distribution, with the continuous ranked probability score (CRPS) as a constraint in the loss function. Lang et al.⁴³ developed the AIFS-CRPS ensemble model as a variant of the deterministic AIFS system, utilizing CRPS as its loss function. Other studies adopted generative techniques, such as the diffusion models, which start with a pattern of random noise and iteratively refine this noise into coherent images. For instance, Price et al.⁴⁴ developed GenCast, a diffusion model that generates ensemble forecasts by sampling from a joint probability distribution of potential weather scenarios across space and time. Similarly, Li et al.⁴⁵ employed a diffusion model in their Scalable Ensemble Envelope Diffusion Sampler (SEEDS), which generates large ensembles based on two members of the Global Ensemble Forecast System at the National Centers for Environmental Prediction (NCEP). These models have shown improved ensemble forecast skill scores in certain metrics compared to the operational ensemble forecasts at ECMWF. However, ensemble members and perturbations in these AI-driven models are generated within hidden layers of neural networks, making it challenging to explicitly manifest and diagnose the characteristics, evolution, and covariances of the ensemble forecast perturbations, thereby limiting a comprehensive evaluation of these AI-driven ensemble prediction systems.

The focus of this study is initial uncertainty and their temporal evolution within AI models. Based on diagnostics of error growth dynamics, we developed an ensemble forecast scheme with one of the most advanced AI models for effective TC track ensemble forecasts. Our innovative attempt superimposes 3D perturbations with flow dependence, physical constraints, and finely tuned magnitudes onto the initial conditions of the AI model for perturbed forecasts. Detailed diagnostics reveal that these initial perturbations, unlike randomly distributed small-amplitude initial perturbations, exhibit reasonable dynamic growth properties within the AI model, similar to those in NWP models. Specifically, the short-range ensemble perturbation covariance produced by our scheme exhibits similarity to those from the ensemble forecasts of ECMWF. As a result, our AI-based TC track ensemble forecasts significantly outperform those of ECMWF in terms of ensemble mean track error, spread-skill ratio (SSR), and CRPS using the same ensemble size of 50. Furthermore, our results demonstrate enhanced TC track forecasts when the ensemble size is increased to 2000, which has never been tested in prior studies.

Results

Perturbation growth in the IFS and FuXi models

It is crucial to understand a model's error growth dynamic before generating appropriate perturbations for it⁴⁶. This holds particular significance for AI weather models, which are derived from data training rather than the explicit application of atmospheric physics laws. To compare the error growth dynamics between the IFS and FuXi models, Fig. 1 illustrates their perturbation growth rates within 72-h lead times as a function of the magnitude of initial perturbations, using the lagged forecast method (see more details in the “Calculation of perturbation growth rate” section). The results for other lead times are qualitatively similar (see Supplementary Fig. 1). This analysis covers three defined domains: the western North Pacific (WNP), the North Atlantic (NA), and the Northern Hemisphere (NH), to provide a more comprehensive understanding. It is noted that the magnitude of initial perturbations is adjusted by setting different lagged intervals of forecasts ($i\Delta t$), leading to differences in initial perturbation magnitudes between the IFS and FuXi models given a certain $i\Delta t$. For the FuXi model, perturbation growth is also analyzed for initial Gaussian-distributed noise,

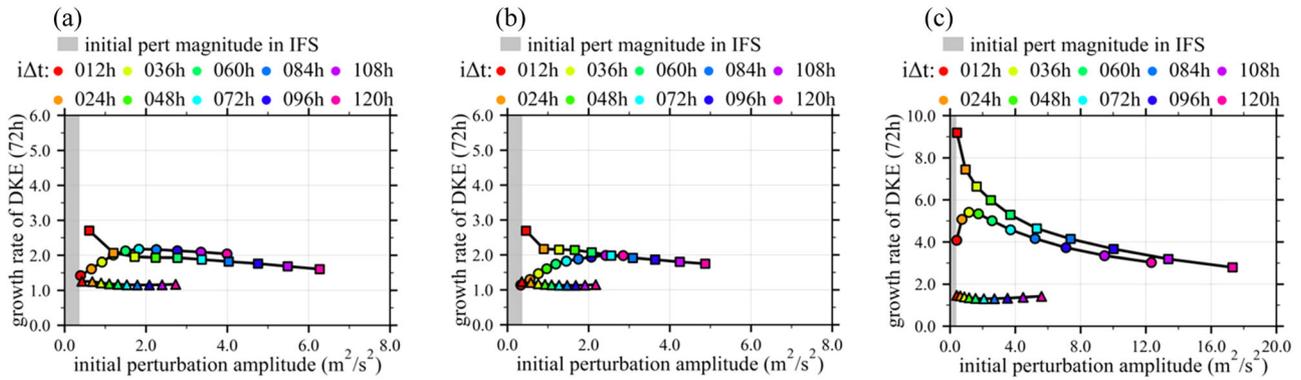


Fig. 1 | The relationship between 72 h growth rate and initial magnitude of deep wind perturbations. Perturbations and growth rates are evaluated in the western North Pacific Ocean (a), the North Atlantic Ocean (b), and the mid-latitudes (c) from July to October 2021. Different colors represent different lagged intervals ($i\Delta t$).

Dots, boxes, and triangles correspond to evolved perturbations in FuXi, evolved perturbations in IFS, and Gaussian noise in FuXi, respectively. The gray area indicates the initial perturbation magnitude in the IFS ensemble forecast.

which has the same statistical mean and standard deviation as the initial evolved perturbations of the lagged forecast method for comparison (refer to the “Calculation of perturbation growth rate” section).

The 72-h perturbation growth rates in WNP, NA, and NH regions for the IFS decrease as the magnitude of initial perturbations increases, indicating a more rapid growth for smaller initial perturbations. This is consistent with previous studies that attribute this phenomenon to more intense atmospheric instability at smaller scales^{47,48}. In contrast, the growth rate of initial evolved perturbations for the FuXi model is significantly lower than that in IFS when the initial perturbation magnitude is below approximately $1.5 \text{ m}^2 \text{ s}^{-2}$. In addition, the growth rate of perturbations in FuXi increases with larger magnitudes, displaying an opposite trend to that observed in the IFS. This discrepancy suggests that the dynamical growth of small perturbations in the AI-based weather model may differ from those in NWP models, and the dynamical sensitivity is much less pronounced in AI models. This finding supports the conclusion by Selz and Craig⁴⁰ that AI weather models may have limitations in simulating the “butterfly effect”.

Although FuXi exhibits significantly weaker growth dynamics with small initial perturbations, our findings reveal that when the magnitude of initial perturbations exceeds certain thresholds (about $1.5\text{--}2 \text{ m}^2 \text{ s}^{-2}$ for the WNP, NA, and NH), the growth rate becomes similar to that of physics-based dynamic models. This threshold is comparable to the amplitude of analysis error variance estimated in most global operational data assimilation systems^{49,50}. This phenomenon may stem from the nature of AI models, which are trained on historical datasets of limited duration and spatial resolution. Since the minimum deviation of initial states between analogous trajectories is constrained in historical datasets, models trained on it have limited ability to resolve the evolution of perturbations with amplitudes below the threshold. In addition, random noise perturbations, despite having the same magnitude as the initial evolved perturbations, exhibit a much slower growth rate (slightly above one) in FuXi. Contrary to previous studies suggesting that the perturbation dynamics in AI models may significantly differ from those in NWP models, our results demonstrate that AI-based models exhibit similar perturbation growth when initial perturbations have appropriate magnitude and physical constraints. This finding quantitatively complements the results from Selz and Craig⁴⁰ and provides a basis for selecting initial perturbation amplitude in ensemble forecasting.

To further assess the physical consistency of perturbation evolution within the FuXi and IFS, Fig. 2 illustrates the 5-day temporal evolution of 500-hPa geopotential height (GH) perturbation fields for both IFS and FuXi, starting from 0000 UTC on July 20, 2021. During this period, TC In-Fa (highlighted by green stars) was developing in WNP. A perturbation with a lagged interval of 36 h (i.e., $i\Delta t = 36 \text{ h}$) is used as the initial perturbation for both FuXi and IFS. The evolution of forecasts with random Gaussian noise in FuXi is also presented for comparison.

Figure 2 clearly shows that the initial evolved perturbations of 500-hPa GH in both FuXi and IFS predominantly display large-scale, wave-like patterns at mid-latitudes. These perturbations are primarily concentrated around established weather patterns, including the troughs at the Scandinavian Peninsula and North America, and the ridge in Siberia, which are associated with synoptic-scale baroclinic instabilities^{51,52}. Although less intense, notable initial perturbations are also evident in the tropics, particularly around the TC vortices and their surrounding environments in the WNP. This indicates that these dynamically evolved initial perturbations can effectively capture uncertainties in initial conditions that are closely related to the instabilities of these weather patterns. In stark contrast, initial perturbations derived from random noise lack this spatial coherence.

As the lead time increases, the initial evolved perturbations of 500-hPa GH in both the FuXi and IFS models gradually develop and amplify, driven by the dynamics of the background flow. Notably, these perturbations demonstrate a high degree of spatial similarity across both models. For instance, perturbations intensify and cluster around the deepening troughs near the eastern coast of North America and the Ural Mountains, as well as around the TC vortices. The kinetic energy spectra of perturbations in both the FuXi and IFS models, as a function of lead time, are also presented to quantify their similarity (see a7 and c7 in Fig. 2). The temporal variation of the spectrum in the evolved perturbations from FuXi and IFS shows a similar pattern, both demonstrating clear upscale perturbation growth behavior. However, the Gaussian noise in FuXi displays significantly higher perturbation energy at small scales and substantially weaker energy at wavelengths above 300 km compared to the reference throughout all lead times. Qualitatively similar results are observed in other cases (see Supplementary Figs. 2–4).

This spatial similarity can be attributed to two primary aspects. Firstly, the FuXi model presents robust capability in simulating large-scale circulation patterns, similar to that of the IFS model. Secondly, the initial perturbations are appropriately constrained in physics and magnitude, allowing FuXi to effectively simulate their dynamical evolution on the reference flow pattern. In contrast, the evolution of random noise exhibits significantly less physical consistency. The growth of Gaussian noise within FuXi is notably slower and less organized compared to the initial evolved perturbations, even with comparable initial perturbation amplitudes, as is also noted by Selz et al⁴⁰. This disparity suggests a potential limitation of FuXi in identifying and adjusting random noise that is rarely observed in the training dataset and appears to project onto decaying modes. Moreover, AI models still have difficulty in accurately representing small-scale motions, resulting in minor peaks at small scales in the FuXi spectrum. This issue has also been discussed in Selz and Craig⁴⁰ and Bonavita⁴¹.

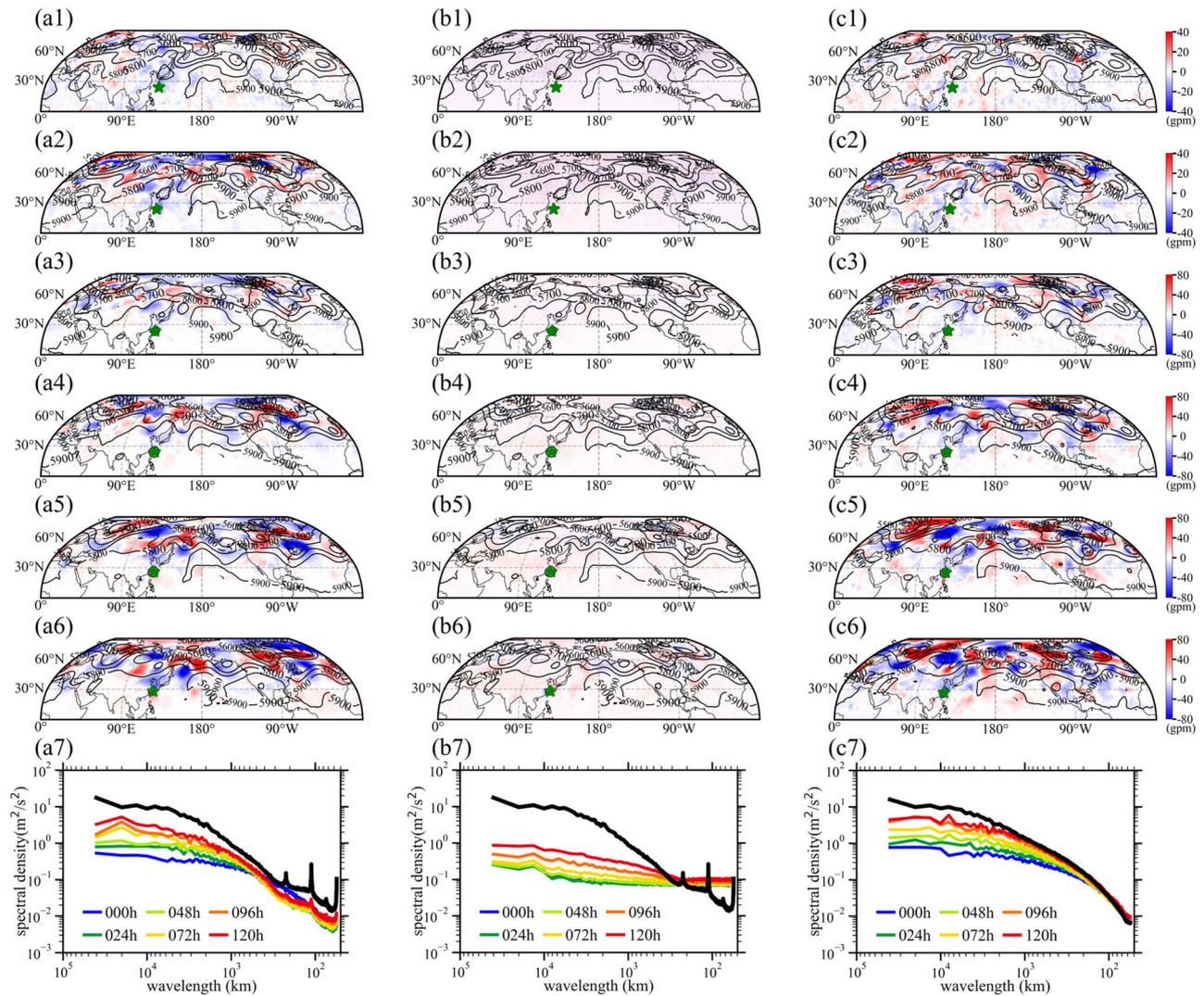


Fig. 2 | The growth of different perturbations in FuXi and IFS for the forecast initialized at 0000 UTC Jul 20, 2021. Columns a–c correspond to perturbations, Gaussian noise in FuXi, and perturbations in IFS, respectively. Rows 1–6 correspond to the forecasts at 0 h, 24 h, 48 h, 72 h, 96 h, and 120 h, respectively; Contours in these figures indicate the 500-hPa geopotential height from the control forecast, colored shading is the differences in 500-hPa geopotential height between the perturbed

forecast and the control forecast. Green stars represent locations of TC In-Fa. Row 7 gives the kinetic energy spectra of 500 hPa wind in control forecast (black line, mean spectra of 0–5-day forecast) and the kinetic energy spectra of 500 hPa wind in evolutionary perturbations (colored lines, colors represent different lead times). The kinetic energy spectra are evaluated between 0N and 70N.

TC track ensemble forecast

Building on the insights from the quantitative analysis on the dynamic growth of perturbations in the FuXi model, this section emphasizes the development of TC ensemble forecasts and the evaluation of TC track forecasts within this model. Initially, the skill of the deterministic forecasts for TC tracks in FuXi is assessed prior to the implementation of ensemble predictions. The results (see Supplementary Fig. 5) show that the average TC track errors over five days are overall lower in the FuXi model compared to the IFS, with reductions from 370 km to 300 km in the WNP and from 410 km to 200 km in the NA. This improvement in TC track forecast accuracy in FuXi aligns with findings in previous studies^{27,28}, laying a solid foundation for subsequent TC track ensemble forecasting efforts.

In the ensemble generation scheme of the FuXi model, the lagged interval of forecasts ($i\Delta t$) is a critical parameter that governs the magnitude of initial perturbations in the ensemble forecasting (refer to “A fast physics-based perturbation generator for the FuXi model”). Sensitivity tests have identified an optimal lagged interval of 36 h, which minimizes the ensemble mean track forecast errors averaged over all samples and

ensures that the SSR of the track error approaches one (see Supplementary Fig. 6). This selection is further supported by quantitative analyses presented in Figs. 1 and 2. As demonstrated in Fig. 1, the initial perturbation amplitude of DKE over the tropics with a 36-h lagged interval in FuXi is approximately $0.8 \text{ m}^2 \text{ s}^{-2}$, which is larger than the initial ensemble spread of the operational IFS ensemble forecast, approximately $0.3 \text{ m}^2 \text{ s}^{-2}$. However, the growth rate of these perturbations is slower than that of the IFS ensemble (1.3 vs. 4.2 for 48 h in WNP). This discrepancy in initial perturbation amplitude can be attributed to differences in the ensemble generation schemes used in IFS and FuXi, as well as their fundamental differences in model dynamics. Specifically, IFS utilizes the singular vector (SV) and ensemble data assimilation to generate initial perturbations. In SV, the fastest linearly growing perturbation conditioned on a specific atmospheric state over a short subsequent time window is represented. The dynamic properties of the SV likely contribute to its accelerated growth in the initial days. Consequently, despite larger perturbation amplitudes and slower growth rates, the 36-h lagged interval is selected due to its comparable perturbation amplitudes and ensemble spread of

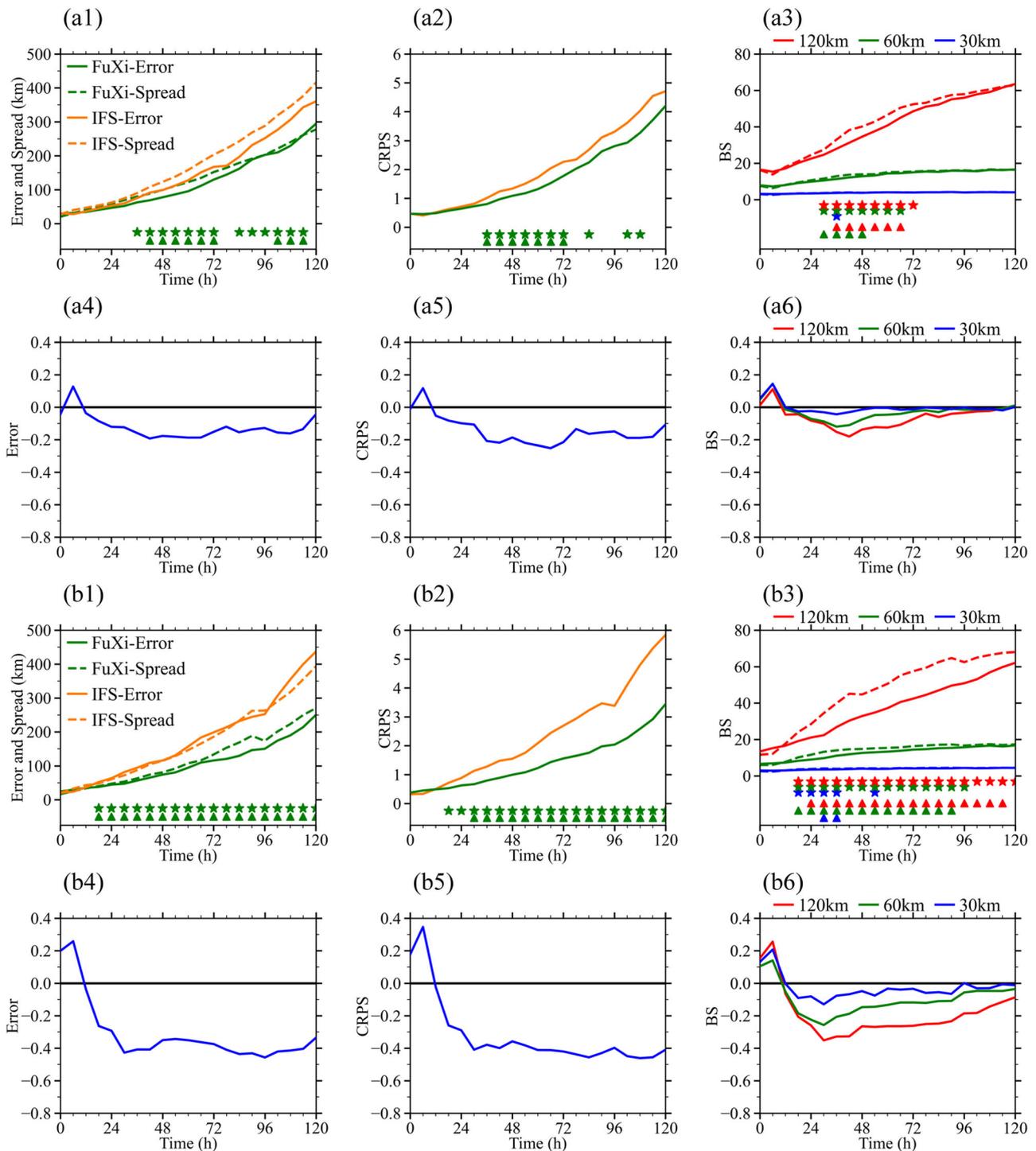


Fig. 3 | Ensemble forecast skills of all 113 typhoon cases. **a1** gives FuXi (green line) and IFS (yellow line) mean track error (solid line) and ensemble dispersion (dashed line). **a2** gives the CRPS scores of the ensemble track forecasts from FuXi (green line) and IFS (yellow line). **a3** gives the BS scores of FuXi (solid line) and IFS (dashed line) track forecasts at thresholds of 30 km (blue), 60 km (green), and 120 km (red). Scores in **(a1–a3)** are averaged over 74 typhoon forecasts in WNP. **a4–a6** give the relative differences in ensemble mean errors, CRPS scores, and BS scores between FuXi and IFS. Negative (resp. positive) error, CRPS, and BS values indicate FuXi (resp. IFS) performs better. Stars (triangles) in each figure show lead times when mean track error, CRPS and BS scores are significantly lower in FuXi than in IFS at a 90% (95%) confidence level with the student *t*-test. **b1–b6** are the same as **(a1–a6)** but correspond to 39 typhoon forecasts in NA.

environmental variables for TCs at 48 and 72 lead times. A longer (shorter) lagged interval results in larger (smaller) amplitudes of initial perturbations, thereby overestimating (underestimating) the ensemble spread of TC tracks (see Supplementary Fig. 6). Figure 2 further reinforces that initial perturbations with a 36-h lagged interval effectively capture the physics-based uncertainties in the initial conditions, thereby providing

reliable dynamical evolution of forecast perturbations in subsequent forecasts.

Figure 3 compares the TC track ensemble forecast skill averaged over all 113 cases between the FuXi and IFS. Both ensemble forecasts utilize a size of 50 members. The assessment metrics include mean track error, SSR, BS, and CRPS (see more details in the “Evaluation metrics” section),

highlighting their relative differences. As depicted in Fig. 3a1, a4, the ensemble mean track error for FuXi is approximately 20% lower than that in the IFS ensemble in the WNP basin beyond day 1. The reduction in track error is even more pronounced in the NA basin, reaching up to 40% (Fig. 3b1, b4). The reduction is statistically significant at the 0.1 level in 2- to 5-day forecasts. This may be attributed to the superior performance of the FuXi deterministic TC track forecast (see Supplementary Fig. 7 for similar error reductions in deterministic forecasts and ensemble mean forecasts). Furthermore, the ensemble spread of TC tracks in FuXi maintains relatively consistent with its ensemble mean errors, indicating a high reliability score for the FuXi ensemble forecast. This consistency is crucial for an EPS, as it ensures that the TC track ensemble members effectively represent forecast uncertainties and accurately reflect the spatial range of TC positions.

In terms of the probability forecast of TC tracks, the FuXi ensemble outperforms the IFS ensemble as shown by the CRPS, with improvements of approximately 20% in the WNP and 40% in the NA. This suggests that the probability density function (PDF) of the TC position ensemble generated by FuXi more accurately reflects the observed probability distribution of TC positions. Additionally, the FuXi ensemble demonstrates superior performance in the BS at thresholds of 120 km, 60 km, and 30 km, as shown in Fig. 3a6, indicating that the FuXi TC track ensemble forecast is more effective in estimating the strike probability at varying thresholds. For CRPS and BS, the reduction is also statistically significant in the forecast from 2 to 3 days. Despite these systematic improvements in TC track ensemble forecasting skill, the FuXi ensemble exhibits a 10%–20% degradation in performance during the first 12 h, as measured by the applied metrics, which might be caused by the larger amplitude of the initial perturbations in the FuXi model. The statistically insignificant differences in the performance of BS and CRPS beyond 4 days might be attributed to the reduced number of cases.

To further evaluate the forecast skill of our method, we compare the evolved perturbation generator with a Gaussian perturbation generator and a generative-based FuXi-ENS ensemble system using 27 forecasts from 5 TC cases in 2018 (TCs Ampil, Ernesto, Cimaron, Florence, and Kong-rey). Initial Gaussian noise is generated from the standard deviation of initial evolved perturbations. Details of the FuXi-ENS scheme and its TC track forecast can be found in Zhong et al.⁴². Notice that only a limited number of TC cases from Zhong et al.⁴² are used in this study for the initial comparison, as FuXi-ENS has not yet been open sourced. A more comprehensive comparison, involving a larger number of TC cases, will be conducted in future studies. The forecast verification results and the ensemble forecast tracks for two TC examples are provided in Supplementary Figs. 8 and 9, respectively. Notably, although FuXi-ENS demonstrates somewhat superior performance in mean track error and CRPS scores, FuXi with the evolved perturbations exhibits higher spread, which maintains the consistency between ensemble spread and mean track error, along with lower BS scores. The spread-error difference is less than 60 km in FuXi with evolved perturbations, comparable to the skills of IFS, whereas the difference exceeds 100 km in FuXi-ENS and 200 km in FuXi with Gaussian noise. As shown in Supplementary Fig. 9, ensemble tracks generated by the evolved perturbation generator exhibit a larger spread and encompass the IBTrACS track, outperforming other methods.

Ensemble forecast with 2000 members

To further evaluate the forecasting capabilities of the FuXi ensemble, particularly its advantage in rapid computation, we conducted a case study on the track forecasts of Typhoon Chanthu using 2000 members, starting from 0000 UTC on September 9, 2021 (see Fig. 4). This study aims to investigate the impact of substantially large ensemble sizes on TC forecasts and evaluate their contribution to the forecast accuracy. Additionally, it addresses the trade-off between the number of ensemble members and the associated computational cost, which was previously impractical with traditional numerical models but is now feasible with fast and accurate AI-based models. In our experiment, the operational IFS ensemble forecasts with 50 members (IFS-50), FuXi ensemble forecasts with 50 members (FuXi-50), and 2000 members (FuXi-2000) were analyzed to assess performance

differences. Ensemble forecast generated with Gaussian noise in FuXi is also tested, as shown in Supplementary Fig. 10. It is noted that Gaussian noise without physically consistent spatial distribution grows slowly in FuXi, leading to a significant underestimation of forecast spread, which is particularly detrimental for reliable probability forecasts. Therefore, the comparisons below are focused on the FuXi ensemble generated with our scheme (FuXi-50 and FuXi-2000) and the ensemble from the operational IFS ensemble (IFS-50).

The IFS-50 and FuXi-50 ensemble forecasts demonstrate comparable performance in predicting the track of Typhoon Chanthu, with both systems achieving a 120-h mean track error of less than 150 km. When evaluated using mean track error, BS, and CRPS, FuXi-50 outperforms within the first 100 h of the forecast, while IFS-50 exhibits greater accuracy in the 100–120 h forecast range, as illustrated in Fig. 4d, h, l. The track spread in the IFS ensemble remains slightly larger throughout the forecast period. For instance, in the 120-h IFS ensemble forecast, 39.53% of the ensemble members deviate by over 240 km from the IBTrACS track, whereas in the FuXi ensemble, only 19.15% of the members exhibit this level of deviation.

The comparison between FuXi-2000 and FuXi-50 clearly demonstrates that increasing the number of ensemble members improves probabilistic scores, such as CRPS and BS, by approximately 10% (Fig. 4l), while it has a minimal effect on the ensemble mean error. In particular, the frequency distribution of TC positional errors becomes more continuous and accurate as the ensemble size increases. For instance, in the IFS-50 and FuXi-50 forecasts, the frequency distribution of ensemble errors exhibits multiple maxima and minima, and the spatial coverage of ensemble tracks is discrete (Fig. 4c, g). In contrast, FuXi-2000 produces a much smoother frequency distribution with a prominent peak around 120 km, and the spatial coverage of ensemble tracks provides a continuous and nuanced representation of error patterns (Fig. 4k). As highlighted in previous research⁵³, increasing the ensemble size allows for the simulation of extreme events at the tails of the forecast distribution, such as $\pm 2\sigma$ and $\pm 3\sigma$, that are not captured in smaller ensembles. For example, in the 120-h forecast by FuXi-2000, some ensemble members predict TC landfall in Japan, an outcome not observed in the FuXi-50 or IFS-50 ensembles (Fig. 4c, g, k). Therefore, using a large ensemble size in ensemble forecasts, particularly with AI models, holds significant potential for enhancing forecasting skills, especially in predicting TC tracks. Similar results are observed in a low predictable TC Haikui (Supplementary Fig. 11). Nonetheless, caution must be taken when extending these single-case results to a broader range of scenarios. The optimal number of ensemble members remains an open question, requiring further theoretical and quantitative analysis.

Comparison of ensemble covariance

To further assess the physical consistency of the FuXi and IFS ensemble forecast, we analyzed the spatial ensemble covariance of 500-hPa GH for FuXi-50, FuXi-2000, and IFS-50, as shown in Fig. 5. In an optimal ensemble prediction scheme, the ensemble should accurately capture the spatially coherent relationships of variables at both the initial and forecast times. In this analysis, the spatial correlation of ensemble perturbations is calculated with respect to the reference point at 30N, 135E (dark green star), which is located near the western boundary of the Western Pacific subtropical high (WPSH).

In Fig. 5, IFS-50 exhibits a reasonable and physically consistent spatial distribution of initial perturbation correlation from the reference point. High levels of positive correlation are observed near the reference point, particularly over the western and northern areas of the WPSH, with a gradual decrease in correlation at greater distances. This pattern can be attributed to the dynamics-based singular vector scheme used to generate IFS-50's initial perturbations^{16,54}. Relatively weak spatial correlations are observed in regions several thousand kilometers away, likely due to the spurious long-distance correlations resulting from the limited number of ensemble members.

The initial perturbation correlation of FuXi-50 is generally similar to that of IFS-50 but exhibits a considerably broader coverage of positive

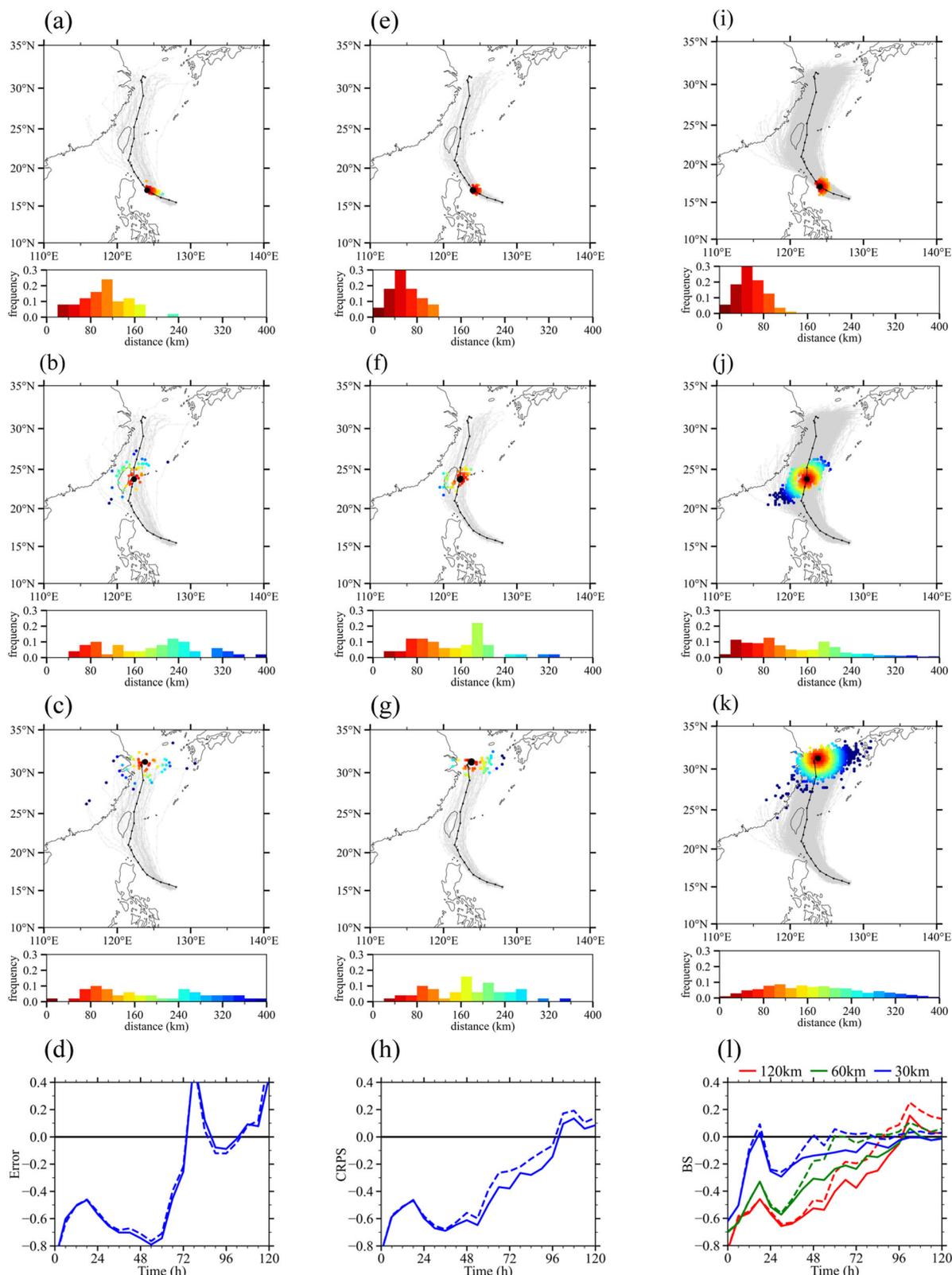


Fig. 4 | The forecast of Chanthu initialized at 0000 UTC Sep 9, 2021 from different models. TC track ensemble forecast from 50 members of the IFS operational ensemble system (IFS-50, a–c), 50 members of FuXi (FuXi-50, e–g), and 2000 members of FuXi (FuXi-2000, i–k) are shown. a–c represent 24 h, 72 h, and 120 h forecast of IFS-50. In the upper subplot, colored dots give TC positions from different forecast members at a valid time, colors indicate forecast errors, and gray lines and black lines give TC tracks from the ensemble forecast and IBTrACS dataset, respectively. TC position of IBTrACS at valid time is marked by a larger black point.

In lower subplots, histograms give the frequency of forecast errors from ensemble forecasting members. e–g and i–k are the same as (a–c), but for FuXi-50 and FuXi-2000, respectively. d shows the relative difference in the ensemble mean track error between IFS and FuXi-50 (dashed line) as well as FuXi-2000 (solid line). h is the same as (d), but for the relative difference in CRPS scores. l is the same as (d) but for the relative differences in BS scores at 30 km (blue), 60 km (green), and 120 km (red) threshold. Negative (resp. positive) values of BS and CRPS indicate FuXi (resp. IFS) performs better.

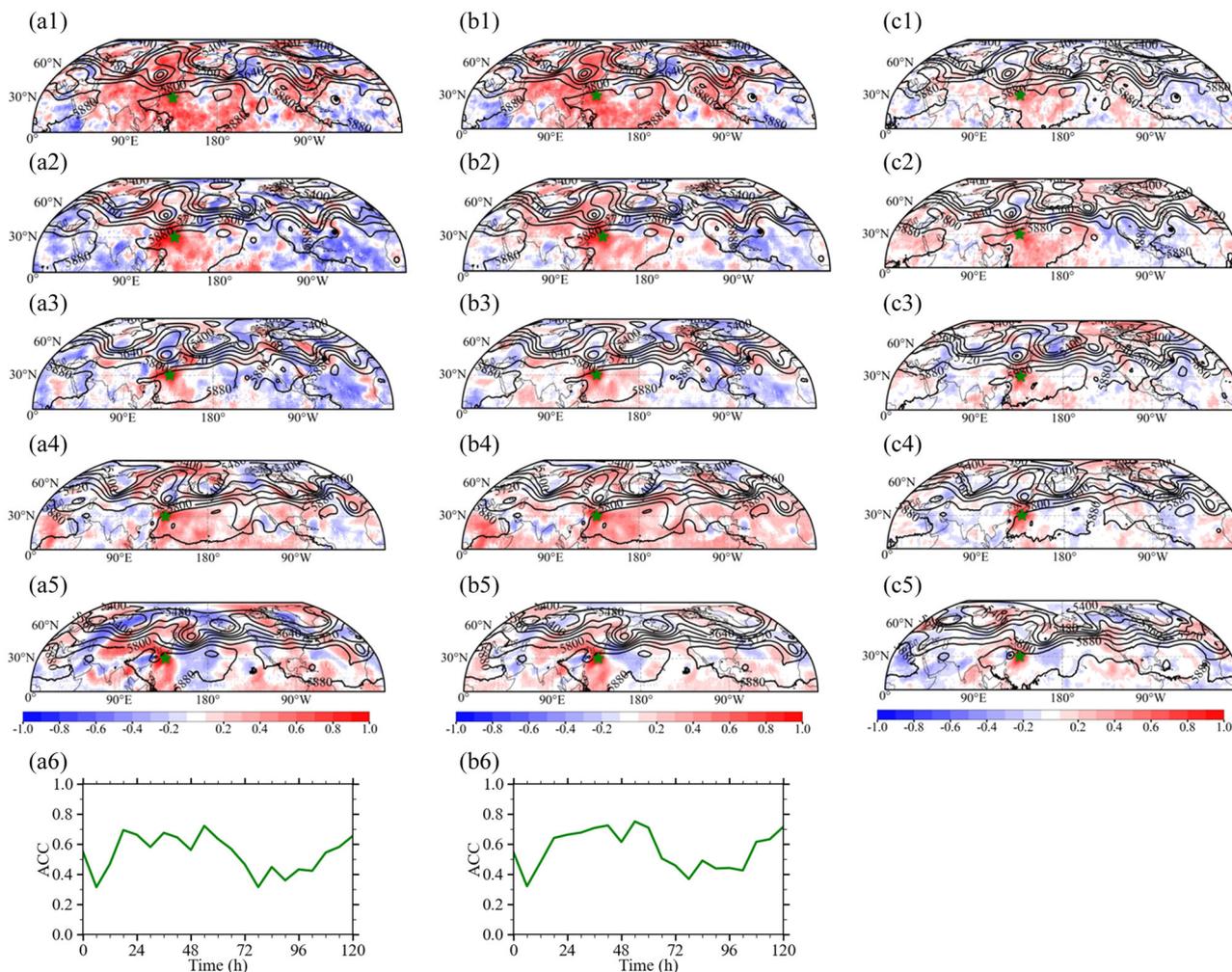


Fig. 5 | Ensemble-based autocorrelations from the green asterisk (30N, 135E) for forecasts from FuXi and IFS. 0 h (row 1), 24 h (row 2), 48 h (row 3), 72 h (row 4), and 120 h (row 5) forecast of 500 hPa geopotential height given by 50 members of FuXi (left column), 2000 members of FuXi (middle column), and 50 members of IFS

(right column) are shown for the forecast of Chanthu initialized at 0000 UTC Sep 9, 2021. **a6** gives the correlations between FuXi-50 and IFS autocorrelations in the WNP area. **b6** is same as **(a6)** but for FuXi-2000.

correlations, extending into the upstream regions in Eastern Europe and Siberia. This broader coverage may result from the generation of initial perturbations of FuXi-50 using a short-time window preceding the initial prediction time. Although FuXi-50 overestimates the area with positive correlations at the initial time, the spatial correlation of the ensemble forecast perturbations in FuXi-50 and IFS-50 becomes more similar as lead time progresses, peaking in 2–3 days (with a spatial correlation coefficient of approximately 0.56 in the WNP basin at 48 h). This similarity may be related to the approximately linear error growth regime within the first two days¹⁶. Furthermore, this similarity remains over 0.5 when using 2000 ensemble members (with spatial correlation coefficient increased to approximately 0.62 at 48 h, Fig. 5b6), indicating that the initial perturbations produced by our physics-based perturbation generation scheme in FuXi develop with physical consistency, akin to those in traditional NWP models. Despite the similarity between the ensemble perturbation covariance of FuXi and IFS ensemble, both exhibit long-distance weak covariance. This may be attributed to the limited dimensionality of the subspace spanned by the perturbations, potentially leading to a rank deficiency of the perturbation covariance matrix.

Beyond 72 h, the similarity between the spatial correlation fields of ensemble perturbations in IFS-50 and FuXi-50 decreases, possibly due to the increasing influence of nonlinear effects. The similarity between IFS-50 and FuXi-50 is also observed for a reference point on land and in mid-latitude

(Supplementary Fig. 12), as well as in TC Haikui (Supplementary Fig. 13). In conclusion, based on selected cases, the similarity in perturbation correlations between FuXi-50, FuXi-2000, and IFS-50 suggests that the perturbations generated by our method in FuXi exhibit physically consistent growth, contributing to improved TC track forecasts. A more comprehensive investigation is needed to fully generalize these results.

Discussion

For high-impact weather systems like TCs, ensemble forecasting is crucial for capturing uncertainty in forecast outcomes to quantify the risk of associated disasters. However, traditional ensemble forecasting of NWP models is limited by the significant computational resources required to run a large number of ensemble members, which hampers further improvements in TC ensemble forecasting skills. The recent development of AI-based weather prediction models, with high forecast accuracy and significantly lower computational costs, presents a promising opportunity for advancing TC ensemble forecasts. Nonetheless, an efficient initial ensemble generation scheme for AI-based models has yet to be developed, that is relevant to the specificities of AI model dynamics. In this study, we first analyze the quantitative relationship between the perturbation amplitude and the growth rate in AI models. It is identified that in the FuXi model, initial perturbations with small amplitude and random distribution grow slowly, while perturbations with appropriate amplitude and physical

constraints exhibit similar dynamical growth behavior as those in NWP models. Based on this, we proposed an efficient initial perturbation generation scheme for the FuXi model, which brought superior ensemble forecast skill for TC tracks compared to the state-of-the-art EPS of ECMWF.

The sensitivity analysis of initial perturbations reveals that the dynamical growth of perturbations in the FuXi model is influenced not only by the amplitude but also by the spatial characteristics. Initial perturbations with small amplitudes in the FuXi model grow significantly slower than those with the same amplitude in the IFS model of ECMWF, consistent with preceding findings⁴⁰. Moreover, initial perturbations consisting of random noise, regardless of the amplitude, exhibit slow and non-physical growth dynamics. However, our study extends beyond prior research by demonstrating the quantitative relationship between the perturbation amplitude and the growth rate, showing that initial evolved perturbations, once they surpass a certain amplitude threshold, exhibit growth rates and spatial structure development comparable to those in the NWP model. This is attributed to the generation of these initial perturbations, which incorporate the short-range evolutionary dynamics inherent to the FuXi model, thereby reflecting the uncertainties associated with atmospheric physical instabilities.

We proposed a fast physics-based ensemble generation scheme based on the initial evolved perturbation for the FuXi model and applied it to the ensemble forecasting of TC tracks. The perturbation covariance matrix is estimated using a set of initial evolved perturbations within a two-week time window preceding the initial time of forecasts. This covariance matrix is then utilized to conveniently generate initial perturbations in any desired quantity, ensuring that the generated ensemble adheres to the statistics of the first (mean) and second (covariance) moments of the perturbations.

Our evaluation based on extensive TC samples from both the WNP and NA basins demonstrates that the FuXi ensembles outperform the IFS ensembles with an equal ensemble size of 50 members in terms of not only the ensemble mean TC track errors but also the SSR as a measure of reliability. This leads to higher probability skill scores of the FuXi ensemble forecasts, such as the CRPS and BS. The superior performance of the FuXi ensembles can be attributed to both the physics-conditioned spatial correlations of ensemble perturbations within the model and the improved deterministic TC track forecasts provided by FuXi. It is also expected that, using appropriate adjustments to the ensemble perturbation covariance, such as the rescaling and spatial localization^{55,56}, the FuXi ensemble has the potential to further improve forecast skill. Moreover, due to the high computational efficiency of the FuXi model, we were able to conduct TC ensemble forecasts with 2000 members — a scale unprecedented in previous studies. Preliminary analysis of two TCs indicates that the increase in ensemble size further enhanced the TC ensemble forecast skill. However, further work is needed to generalize the results and better understand the ensemble size required. And characteristics of the error distribution in large ensembles should also be further analyzed.

Although the FuXi model demonstrates improved ensemble forecast skills for TC tracks compared to the IFS, there is still potential to enhance the initial perturbation scheme. In our method, the growing perturbations are not adequately sampled, a degradation in forecast performance exists during the first 12 h, and the flow dependency within initial ensemble perturbations is relatively weak due to the design of the time window. Future research could explore the integration of optimally growing modes, such as the singular vector¹⁶ and the conditional nonlinear optimal perturbation (CNOP⁵⁷), into the perturbation generation, which may lead to further enhancements of TC ensemble forecast skill. On the other hand, notice that this study focuses solely on initial uncertainty by perturbing only the initial states. Representing model deficiencies in ensemble forecasting, particularly using ML models, remains a significant challenge^{39,53,58}, and will be explored in future studies. Moreover, TC intensity, an area where AI models are known to be less skillful, is another crucial aspect to explore through ensemble forecasting in future works. Additionally, while AI-based weather models have limitations, particularly in smoothing small-scale features in longer-range forecasts (WeatherBench 2²⁶), they excel in simulating large-

scale weather patterns, such as TC-related environmental circulation. This capability is crucial for efficient TC track ensemble forecasts using AI models. Therefore, it is essential to gain a comprehensive understanding of the physical and dynamical properties of AI models before developing their effective applications in research and prediction.

Methods

FuXi model

This study utilized FuXi, an AI-based global medium-range weather forecasting model, to investigate the perturbation growth dynamics and develop an effective ensemble forecasting scheme for TC tracks. FuXi model is distinguished by its superior performance in deterministic forecast RMSE and anomaly correlation coefficient (ACC) relative to ECMWF high-resolution IFS and other AI-based models according to the evaluation of WeatherBench 2²⁶.

FuXi model employs the space-time cube-embedding technique to reduce the dimension of the input multi-variable 3D data. The embedded data is then processed using the computationally efficient U-transformer architecture. The model was trained on 39 years of 6-hourly ECMWF ERA5 reanalysis data with a spatial resolution of 0.25 degrees. Predictions are generated through a simple fully connected (FC) layer, producing meteorological variables including five upper-air atmospheric variables (geopotential, temperature, horizontal wind, and relative humidity) and five surface variables (10-m horizontal wind, 2-m temperature, mean sea-level pressure, and 6-hourly total precipitation), all at a horizontal resolution of 0.25 degree. As an autoregressive model, FuXi initiates predictions using two preceding states at times t and $t - 1$ and iterates with a time step of 6 h. For 15-day forecasts, three models fine-tuned from the pre-trained FuXi model, i.e., FuXi-Short, FuXi-Medium, and FuXi-Long, are employed for 0–5 days, 5–10 days, and 10–15 days forecasts, respectively. This cascade architecture with multiple fine-tuned models aims to minimize error accumulation from iterative predictions and optimize performance across different forecast lead times. In this research, as most operational forecast systems^{15,20,59}, we focus on the TC track forecast within 5 days, and thus only the FuXi-Short model is used. Further details on the FuXi model can be found in Chen et al.²⁹.

Forecast and verification data

The primary aim of this study is to compare the TC track ensemble forecasts derived from the FuXi model to those generated by the operational EPS of the ECMWF. Previous research has consistently demonstrated that the ECMWF EPS provides the most accurate TC track forecasts among advanced NWP-based EPSs^{14,15}. The historical operational ensemble forecast products used in this study are sourced from the THORPEX International Grand Global Ensemble (TIGGE) dataset, which archives operational ensemble forecast data from 13 global NWP centers, commencing in October 2006 (available at <https://apps.ecmwf.int/datasets/data/tigge/levtype=sfc/type=cf/>). In addition to the ensemble forecast data that are used to evaluate the TC track probability forecast skill and the related evolution of ensemble covariances, the deterministic (or unperturbed) forecasts of ECMWF are also utilized to explore the dynamical growth properties of perturbations (see more details in the “Calculation of perturbation growth rate” section). Both the ensemble and control forecasts are produced with the IFS version 48r1, configured with a horizontal resolution of 9 km and 137 vertical levels. However, the ECMWF forecast data accessible in the TIGGE archive has been interpolated to a horizontal resolution of 0.25 degrees, facilitating a straightforward comparison with the output data from FuXi.

In addition to conventional prognostic variables, the TIGGE provides the dataset of TC track ensemble forecasts for a direct evaluation of track forecast skill. In ECMWF, the positions of TC centers within these ensemble outputs are determined using a TC tracker algorithm. The tracker identifies the local minimums of mean sea-level pressure (MSLP) and then determines the TC center from these minimums with a cyclonic signature and a warm core⁶⁰. The TC tracker for FuXi is a modified version of the one used by the ECMWF. Key modifications include the incorporation of the local

vorticity maxima as possible TC centers and the application of average-pooling during the detection of local maxima and minima. In which, the average-pooling technology is used to calculate the average value of small regions to avoid false minima. The modified tracker is better suited for FuXi forecasts, which tend to be smoother and less physically constricted and might include false local minima due to minor differences among adjoint points. A comprehensive description of the TC tracker can be referred to Zhong et al.⁴².

We initialize the FuXi model using ERA5 reanalysis data, aligning with standard practices employed by other AI-based weather models^{27,29}. Existing research has shown that AI modes trained on ERA5 also work well with IFS analyses⁶¹. And the ERA5 dataset serves as a reference for verifying forecasts generated by both the ECMWF and the FuXi models. For the evaluation of TC track ensemble forecast skills, the International Best Track Archive for Climate Stewardship (IBTrACS, available at <https://www.ncei.noaa.gov/products/international-best-track-archive>) is utilized as the reference dataset. IBTrACS merges both recent and historical tropical cyclone data from multiple agencies, representing the most comprehensive global collection of TC information available.

Calculation of perturbation growth rate

Given the limited research on the perturbation sensitivity of AI models, we adopted a classical method, i.e., the lagged forecast method as proposed by Lorenz (1982)⁶² to address this issue. In this approach, we consider two forecasts $F_{t_{-i},(j+i)\Delta t}$ and $F_{t_0,j\Delta t}$ initialized at times t_{-i} and t_0 , respectively, with a lagged interval of $i\Delta t$, but both valid at the same time. Their forecast leading times are $(j+i)\Delta t$ and $j\Delta t$, respectively, and Δt represents a time interval. We define the difference between the two forecasts as $e(i\Delta t, j\Delta t)$, and calculate the mean growth rate (or the amplification factor) of this difference over the leading time $j\Delta t$ as follows:

$$r_{i\Delta t, j\Delta t} = \frac{|e(i\Delta t, j\Delta t)|}{|e(i\Delta t, 0)|} \quad (1)$$

where $|\cdot|$ denotes the L_2 norm. In this method, $e(i\Delta t, 0)$ and $e(i\Delta t, j\Delta t)$ represent the initial perturbation and its evolution after a time of $j\Delta t$. It is noted that while analyzing the variation in growth rate in Fig. 1, we use the accumulative kinetic energy of deep-layer wind from 850hPa to 200hPa within local regions as the variable since it best represents TC motion. However, in ensemble forecasts, initial perturbations are generated and applied globally to all input variables, including three-dimensional zonal and meridional winds, temperature, geopotential, and relative humidity.

This method offers several distinct advantages. (1) It can be easily implemented as long as historical forecast data at regular intervals are available, requiring minimal computational resources. (2) The initial perturbation (i.e., $e(i\Delta t, 0)$) is evolved by the model dynamics during the short-range lagged forecast (termed as initial evolved perturbation hereafter), resulting in dynamical perturbations that exhibit physical balance and spatial instability characteristics of the atmosphere. (3) By adjusting the lagged time (i.e., $i\Delta t$), the method allows for flexible control over the magnitude of initial perturbations. These properties make it an effective approach for analyzing the perturbation growth properties of a dynamical system. This method will be employed to compare the perturbation growth rate and its dependence on the magnitude of initial perturbations between the IFS of ECMWF and the FuXi model.

A fast physics-based perturbation generator for the FuXi model

The study on the perturbation dynamics of the FuXi model demonstrates that, when the magnitude is properly tuned, the initial evolved perturbations exhibit a growth rate comparable to those in the ECMWF physics-based model (see Fig. 1). This outcome is closely linked to the generation of initial perturbation, which is calculated as the difference between the FuXi forecast at a lagged interval of $i\Delta t$ and the analysis at the same valid time t_0 . The short-term (i.e., $i\Delta t$) evolution of the FuXi model dynamics progressively drives the perturbation $e(i\Delta t, 0)$ towards the unstable perturbations^{17,63},

effectively reflecting the spatial uncertainties of variables conditioned on the atmospheric state at t_0 . Such physics-based initial evolved perturbations ensure dynamically and physically reasonable evolution and amplification of perturbations in subsequent lead times, mirroring the behavior of the NWP model (see Fig. 2). Consequently, inspired by the NMC method for 3D-Variational data assimilation, we proposed a fast physics-based scheme for generating initial ensemble perturbations in the FuXi model, consisting of three key steps (see the schematic diagram Fig. 6).

Step 1. At the initial time t_0 for TC prediction, perturbations are selected as $e(i\Delta t, 0)|_{t_0}, e(i\Delta t, 0)|_{t_{-1}}, \dots, e(i\Delta t, 0)|_{t_{-n}}$, where the subscript t_{-n} indicates that the evolved perturbation is valid at time $t_0 - n\Delta t$. In other words, these successive evolved initial perturbations are sampled within a short-time window preceding from $t_0 - n\Delta t$ to t_0 with a sampling interval of Δt . In this study, perturbations $e(i\Delta t, 0)$ are calculated every 12 h ($\Delta t = 12h$), and a total of 25 initial perturbations are selected ($n = 24$), i.e., the time window is 12 days before the initial time t_0 . The evolving interval is set to 36 h, i.e., $i = 3$.

Step 2. 25 initial perturbations $e(i\Delta t, 0)$ are positively and negatively paired to make 50 ensemble members, the same as the ensemble size in IFS. These paired members are used to estimate the perturbation covariance matrix, providing a statistical estimation of the uncertainties in the initial conditions. Notably, no localization method is applied in this study to avoid introducing additional adjustable parameters.

Step 3. Once the perturbation covariance matrix is derived, use it to generate initial ensemble perturbations that approximately represent the first and second moments of the initial condition errors. These perturbations are added to the ERA5 reanalysis state to compute perturbed initial states and carry out ensemble forecasts.

The ensemble initialization scheme offers four advantages. (a) The initial evolved ensemble perturbations not only sample the grid-point uncertainties of the initial error but also capture the spatially coherent structures of these errors, thereby maintaining physical consistency. (b) The time window for generating initial perturbations advances with the initial prediction time, ensuring a certain level of flow dependence of initial perturbations. (c) The scheme allows for convenient adjustment of the perturbation magnitude by tuning the lagged interval of forecasts ($i\Delta t$), for an optimal ensemble forecast performance. (d) Initial ensemble perturbations are generated through random sampling based on the perturbation covariance statistics. This ensures both the randomness and equal probability of the initial members while enabling the rapid creation of a large ensemble size.

Experimental design

This study consists of two experimental components. The first one involves analyzing and comparing the dynamics of perturbation growth between the IFS model of ECMWF and the FuXi model. Based on findings from this analysis, the second component focuses on optimizing the ensemble generation scheme for the FuXi model and applying it to TC track ensemble forecasts. The performance of these TC track ensemble forecasts is comprehensively evaluated and compared with the operational EPS of ECMWF.

In the first part, the perturbation growth rates with varying initial perturbation magnitudes are compared between the IFS model and the FuXi model using the lagged forecast approach. These lagged forecasts are derived from the respective deterministic forecasts of the two models, spanning all forecasts from 1 July 2021 to 31 October 2021, at 12-h intervals. The initial perturbation magnitude is adjusted by varying the lagged interval of forecasts, i.e., $i\Delta t$, from 12 h to 120 h in 12-h increments. For the FuXi model, the dynamical growth of Gaussian random noise is also investigated and compared. Gaussian noise is generated through random sampling based on the statistical mean and standard deviation of initial perturbations in the lagged forecast experiments of FuXi. Since the TC track is primarily influenced by the environmental deep steering flow^{3,4}, the kinetic energy of deep wind from 850 hPa to 200 hPa is used as a norm for calculating the dynamical growth of perturbations. Diagnostic analyses are conducted over two regions where TCs frequently occur: the western North Pacific (WNP,

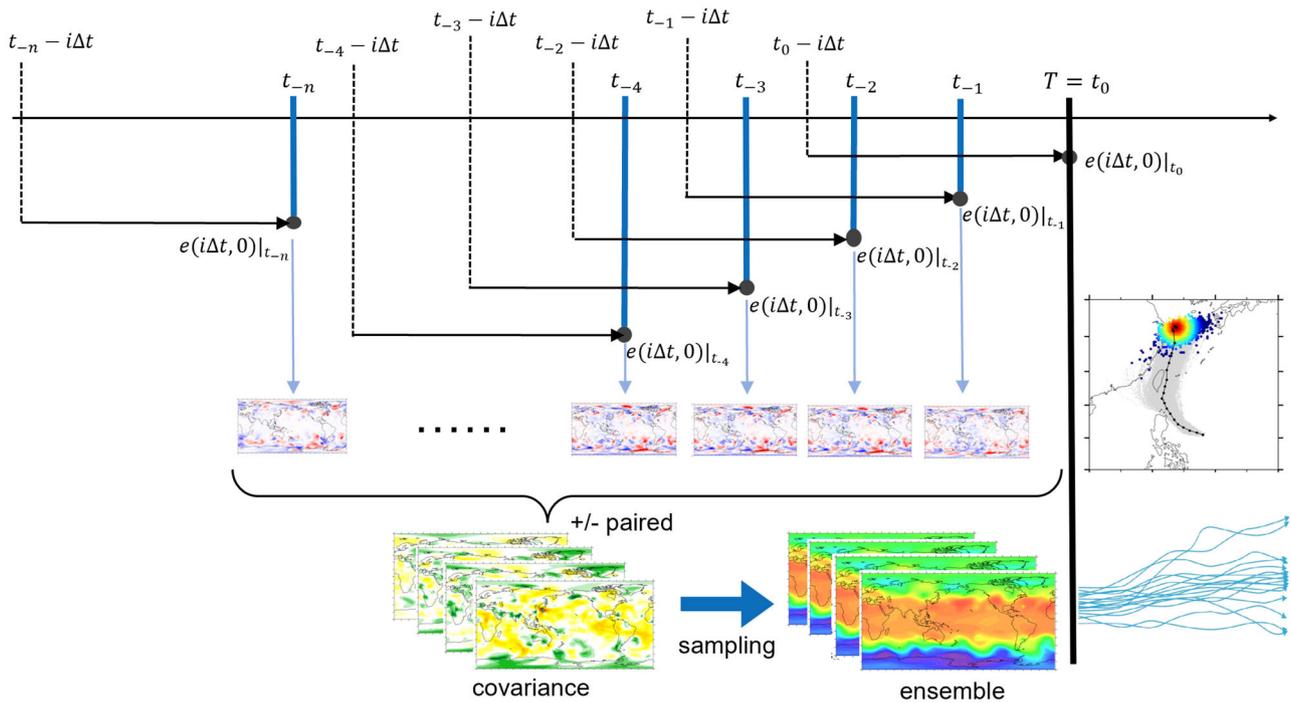


Fig. 6 | Schematic diagram of the initial evolved perturbation scheme. The scheme can be mainly divided into 3 parts: Firstly, 25 evolved perturbations are selected from a time window before the initial time. Secondly, these perturbations are positively

and negatively paired to estimate the perturbation covariance. Finally, initial ensemble perturbations are generated with the perturbation covariance.

100–180E, 0–30N) and the North Atlantic (NA, 20–100W, 0–30N). A broader area covering the mid-latitudes of the Northern Hemisphere (NH, 30–70N) is also included in the analysis.

Following the optimization of the ensemble generation scheme for the FuXi model, its TC track ensemble forecast skill is comprehensively evaluated and compared with those of the IFS ensemble forecasts over many TC samples. The study focuses on TCs in the WNP and NA basins between July and October from 2021 to 2023, selecting cases with a strength above tropical storm level that makes landfall. In total, 36 TCs are selected for analysis. According to the Saffir-Simpson Hurricane Scale⁶⁴, 12 of them are tropical storms; 5 are category 1; 1 is category 2; 3 are category 3; 10 are category 4; 5 are category 5. For each TC, forecast experiments are conducted every two days following the first appearance of the TC in the TIGGE forecast data, resulting in a total of 113 TC ensemble forecast experiments. We select TCs with landfall every two days for the evaluation considering the computational resources. By testing within a subset of TC cases, it is shown in Supplementary Fig. 14 that conducting daily forecast evaluations using all TCs results in only marginal differences in the overall outcomes.

Evaluation metrics

Mean track error and spread are commonly used metrics when evaluating the ensemble forecast skills of TC tracks. They are defined as below:

$$\text{Mean Track Error} = |\bar{p}_k - p_o|$$

$$\text{Spread} = \sqrt{\frac{1}{N-1} \sum_{k=1}^N |p_k - \bar{p}_k|^2}$$

where p_k is the TC location given in latitude and longitude from ensemble member k , p_o is the TC location from IBTrACS, \bar{p}_k is the ensemble mean TC location, N gives the ensemble size and $|\cdot|$ denotes the great-circle distance. A perfect ensemble forecast is characterized by a consistency between spread and error.

CRPS is used to measure the distance between the forecasted TC distribution and the observed TC distribution⁶⁵. It is generalized from one-dimensional CRPS and is as known as energy distance. the CRPS for d -dimensional forecast variable X and observation data Y is defined as:

$$\text{CRPS} = 2E\|X - Y\|_d - E\|X - X'\|_d - E\|Y - Y'\|_d$$

In our research, X is the 2-d TC location from FuXi forecast, Y is the TC location from IBTrACS data, X' and Y' are independent and identically distributed copies of X and Y . E represent mathematical expectation and $\|\cdot\|_d$ represent d -dimensional Euclidean norm. It is shown that lower CRPS indicates a lower difference between FuXi forecasts and IBTrACS data, reflecting better forecast skill⁶⁶. Fair CRPS is employed in this study to compare experimental results.

BS shows the total difference in TC strike probability between FuXi forecast ϕ_f and IBTrACS ϕ_o in region D (WNP or NA in our study), as is also used by Zhang et al.³⁶:

$$\text{BS} = \sum_{p \in D} (\phi_f(p) - \phi_o(p))^2$$

where TC strike probability $\phi(p)$ at location p is defined as the probability that a TC passes within a specific distance, chosen as 120 km, 60 km, and 30 km in our study. A lower BS value gives better forecast skills.

Data availability

ERA5 data is available at <https://cds.climate.copernicus.eu/cdsapp#!/search?type=dataset>; TIGGE data is available at <https://apps.ecmwf.int/datasets/data/tigge/levtype=sfc/type=cf/>; TIGGE Model Tropical Cyclone Track Data is downloaded from <https://rda.ucar.edu/datasets/d330003/>; IBTrACS dataset is at <https://www.ncei.noaa.gov/products/international-best-track-archive>; FuXi model is open sourced at <https://github.com/tpys/FuXi>.

Received: 10 October 2024; Accepted: 14 March 2025;

Published online: 29 March 2025

References

- Peduzzi, P. et al. Global trends in tropical cyclone risk. *Nat. Clim. Chang.* **2**, 289–294 (2012).
- Klotzbach, P. J. et al. Trends in global tropical cyclone activity: 1990–2021. *Geophys. Res. Lett.* **49**, e2021GL095774 (2022).
- Torn, R. D., Elless, T. J., Papin, P. P. & Davis, C. A. Tropical cyclone track sensitivity in deformation steering flow. *Mon. Weather Rev.* **146**, 3183–3201 (2018).
- George, J. E. & Gray, W. M. Tropical cyclone motion and surrounding parameter relationships. *J. Appl. Meteorol. Climatol.* **15**, 1252–1264 (1976).
- Holland, G. J. Tropical cyclone motion: environmental interaction plus a beta effect. *J. Atmos. Sci.* **40**, 328–342 (1983).
- Wu, L. & Wang, B. A potential vorticity tendency diagnostic approach for tropical cyclone motion. *Mon. Weather Rev.* **128**, 1899–1911 (2000).
- Katsube, K. & Inatsu, M. Response of tropical cyclone tracks to sea surface temperature in the Western North Pacific. *J. Clim.* **29**, 1955–1975 (2016).
- Emanuel, K. A. An air-sea interaction theory for tropical cyclones. Part I: steady-state maintenance. *J. Atmos. Sci.* **43**, 585–605 (1986).
- Wang, Y. & Holland, G. J. Tropical cyclone motion and evolution in vertical shear. *J. Atmos. Sci.* **53**, 3313–3332 (1996).
- Lorenz, E. N. The predictability of a flow which possesses many scales of motion. *Tellus* **21**, 289–307 (1969).
- Plu, M. A new assessment of the predictability of tropical cyclone tracks. *Mon. Weather Rev.* **139**, 3600–3608 (2011).
- Kalnay, E. *Atmospheric Modeling, Data Assimilation and Predictability* (Cambridge Univ. Press, New York, 2003).
- Yamaguchi, M. & Majumdar, S. J. Using TIGGE data to diagnose initial perturbations and their growth for tropical cyclone ensemble forecasts. *Mon. Weather Rev.* **138**, 3634–3655 (2010).
- Heming, J. T. et al. Review of recent progress in tropical cyclone track forecasting and expression of uncertainties. *Trop. Cyclone Res. Rev.* **8**, 181–218 (2019).
- Conroy, A. et al. Track forecast: operational capability and new techniques - summary from the Tenth International Workshop on Tropical Cyclones (IWTC-10). *Trop. Cyclone Res. Rev.* **12**, 64–80 (2023).
- Molteni, F., Buizza, R., Palmer, T. N. & Petroliagis, T. The ECMWF ensemble prediction system: methodology and validation. *Q. J. R. Meteorol. Soc.* **122**, 73–119 (1996).
- Toth, Z. & Kalnay, E. Ensemble forecasting at NMC: the generation of perturbations. *Bull. Am. Meteorol. Soc.* **74**, 2317–2330 (1993).
- Zhou, X. et al. Performance of the new NCEP global ensemble forecast system in a parallel experiment. *Weather Forecast.* **32**, 1989–2004 (2017).
- Zhang, X. A. GRAPES-based mesoscale ensemble prediction system for tropical cyclone forecasting: configuration and performance. *Q. J. R. Meteorol. Soc.* **144**, 478–498 (2018).
- Zhou, F. & Toth, Z. On the prospects for improved tropical cyclone track forecasts. *Bull. Am. Meteorol. Soc.* **101**, E2058–E2077 (2020).
- Landsea, C. W. & Cangialosi, J. P. Have we reached the limits of predictability for tropical cyclone track forecasting? *Bull. Am. Meteorol. Soc.* **99**, 2237–2243 (2018).
- Buizza, R. Introduction to the special issue on “25 years of ensemble forecasting. *Q. J. R. Meteorol. Soc.* **145**, 1–11 (2019).
- Leutbecher, M. Ensemble size: how suboptimal is less than infinity? *Q. J. R. Meteorol. Soc.* **145**, 107–128 (2019).
- Irrgang, C. et al. Will artificial intelligence supersede Earth system and climate models? *Nat. Mach. Intell.* **3**, 667–674 (2021).
- Schultz, M. G. et al. Can deep learning beat numerical weather prediction? *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **379**, 20200097 (2021).
- Rasp, S. et al. WeatherBench 2: A Benchmark for the Next Generation of Data-Driven Global Weather Models. *J. Adv. Model. Earth Syst.* **16**, e2023MS004019 (2024).
- Bi, K. et al. Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **619**, 533–538 (2023).
- Lam, R. et al. Learning skillful medium-range global weather forecasting. *Science* **382**, 1416–1421 (2023).
- Chen, L. et al. FuXi: a cascade machine learning forecasting system for 15-day global weather forecast. *npj Clim. Atmos. Sci.* **6**, 1–11 (2023).
- Chen, K. et al. FengWu: pushing the skillful global medium-range weather forecast beyond 10 days lead. Preprint at <https://doi.org/10.48550/arXiv.2304.02948> (2023).
- Leith, C. E. Theoretical skill of Monte Carlo forecasts. *Mon. Weather Rev.* **102**, 409–418 (1974).
- Feng, J., Toth, Z., Zhang, J. & Peña, M. Ensemble forecasting: a foray of dynamics into the realm of statistics. *Q. J. R. Meteorol. Soc.* **150**, 2537–2560 (2024).
- Leutbecher, M. & Palmer, T. N. Ensemble forecasting. *J. Comput. Phys.* **227**, 3515–3539 (2008).
- Buizza, R. et al. A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Weather Rev.* **133**, 1076–1097 (2005).
- Huo, Z. & Duan, W. The application of the orthogonal conditional nonlinear optimal perturbations method to typhoon track ensemble forecasts. *Sci. China Earth Sci.* **62**, 376–388 (2019).
- Zhang, H., Duan, W. & Zhang, Y. Using the orthogonal conditional nonlinear optimal perturbations approach to address the uncertainties of tropical cyclone track forecasts generated by the WRF model. *Weather Forecast.* **38**, 1907–1933 (2023).
- Brenowitz, N. D. et al. A Practical probabilistic benchmark for AI weather models. Preprint at <https://doi.org/10.48550/arXiv.2401.15305> (2024).
- Scher, S. & Messori, G. Ensemble methods for neural network-based weather forecasts. *J. Adv. Model. Earth Syst.* **13**, e2020MS002331 (2021).
- Bülte, C., Horat, N., Quinting, J. & Lerch, S. Uncertainty quantification for data-driven weather models. Preprint at <https://doi.org/10.48550/arXiv.2403.13458> (2024).
- Selz, T. & Craig, G. C. Can artificial intelligence-based weather prediction models simulate the butterfly effect? *Geophys. Res. Lett.* **50**, e2023GL105747 (2023).
- Bonavita, M. On some limitations of current machine learning weather prediction models. *Geophys. Res. Lett.* **51**, e2023GL107377 (2024).
- Zhong, X., Chen, L., Li, H., Feng, J. & Lu, B. FuXi-ENS: a machine learning model for medium-range ensemble weather forecasting. Preprint at <https://doi.org/10.48550/arXiv.2405.05925> (2024).
- Lang, S. et al. AIFS-CRPS: ensemble forecasting using a model trained with a loss function based on the continuous ranked probability score. Preprint at <https://doi.org/10.48550/arXiv.2412.15832> (2024).
- Price, I. et al. Probabilistic weather forecasting with machine learning. *Nature* <https://doi.org/10.1038/s41586-024-08252-9> (2024).
- Li, L., Carver, R., Lopez-Gomez, I., Sha, F. & Anderson, J. Generative emulation of weather forecast ensembles with diffusion models. *Sci Advan.* **10**, eadk4489 (2024).
- Mu, M. & Duan, W. A nonlinear theory and technology for reducing the uncertainty of high-impact ocean-atmosphere event prediction. *Adv. Atmos. Sci.* <https://doi.org/10.1007/s00376-025-4467-9> (2025).
- Ke, J., Mu, M. & Fang, X. Impact of optimally growing initial errors on the mesoscale predictability of heavy precipitation events along the Mei-Yu Front in China. *Mon. Weather Rev.* **150**, 2399–2421 (2022).
- Judt, F. Atmospheric predictability of the tropics, middle latitudes, and polar regions explored through global storm-resolving simulations. *J. Atmos. Sci.* **77**, 257–276 (2020).

49. Feng, J., Toth, Z., Peña, M. & Zhang, J. Partition of analysis and forecast error variance into growing and decaying components. *Q. J. R. Meteorol. Soc.* **146**, 1302–1321 (2020).
50. Feng, J., Wang, J., Dai, G., Zhou, F. & Duan, W. Spatiotemporal estimation of analysis errors in the operational global data assimilation system at the China Meteorological Administration using a modified SAFE method. *Q. J. R. Meteorol. Soc.* **149**, 2301–2319 (2023).
51. Zhang, F. et al. What is the predictability limit of midlatitude weather? *J. Atmos. Sci.* **76**, 1077–1091 (2019).
52. Zhang, F., Bei, N., Rotunno, R., Snyder, C. & Epifanio, C. C. Mesoscale predictability of moist baroclinic waves: convection-permitting experiments and multistage error growth dynamics. *J. Atmos. Sci.* **64**, 3579–3594 (2007).
53. Mahesh, A. et al. Huge ensembles part II: properties of a huge ensemble of hindcasts generated with spherical Fourier neural operators. Preprint at <https://doi.org/10.48550/arXiv.2408.01581> (2024).
54. ECMWF. IFS Documentation CY47R3 - Part V Ensemble prediction system. In *IFS Documentation CY47R3* (ECMWF, 2020).
55. Anderson, J. & Lei, L. Empirical localization of observation impact in ensemble Kalman filters. *Mon. Weather Rev.* **141**, 4140–4153 (2013).
56. Huang, B., Wang, X., Kleist, D. T. & Lei, T. A simultaneous multiscale data assimilation using scale-dependent localization in GSI-based hybrid 4D-EnVar for NCEP FV3-based GFS. *Mon. Weather Rev.* **149**, 479–501 (2021).
57. Mu, M., Duan, W. S. & Wang, B. Conditional nonlinear optimal perturbation and its applications. *Nonlinear Process. Geophys.* **10**, 493–501 (2003).
58. Baño-Medina, J., Sengupta, A., Watson-Parris, D., Hu, W. & Monache, L. D. Towards calibrated ensembles of neural weather model forecasts. Preprint at <https://doi.org/10.22541/essoar.171536034.43833039/v1> (2024).
59. Alaka, G. J., Zhang, X. & Gopalakrishnan, S. G. High-definition hurricanes: improving forecasts with storm-following nests. <https://doi.org/10.1175/BAMS-D-20-0134.1> (2022).
60. ECMWF. Newsletter No. 102 - Winter 2004/05. *ECMWF Newsllett.* (2005).
61. Ben-Bouallegue, Z. et al. The Rise of Data-Driven Weather Forecasting: A First Statistical Assessment of Machine Learning-Based Weather Forecasts in an Operational-Like Context. *Bull. Amer. Meteor. Soc.* **105**, E864–E883 (2024).
62. Lorenz, E. N. Atmospheric predictability experiments with a large numerical model. *Tellus* **34**, 505–513 (1982).
63. Legras, B. & Vautard, R. A guide to Liapunov vectors. ECMWF. In *Proceedings 1995 ECMWF Seminar on Predictability*. Vol. 1, 143–156 (1996).
64. Knapp, K. R., Kruk, M. C., Levinson, D. H., Diamond, H. J. & Neumann, C. J. The International Best Track Archive for Climate Stewardship (IBTrACS): unifying tropical cyclone data. *Bull. Am. Meteorol. Soc.* **91**, 363–376 (2010).
65. Gneiting, T. & Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **102**, 359–378 (2007).
66. Székely, G. J. & Rizzo, M. L. Energy statistics: a class of statistics based on distances. *J. Stat. Plan. Inference* **143**, 1249–1272 (2013).

Acknowledgements

We sincerely thank Drs. Yuejian Zhu, Wansuo Duan, Jiping Guan, Jing Chen, Lili Lei, Xubin Zhang, and Ruiqiang Ding for their valuable comments on this study. The constructive comments of two anonymous reviewers greatly improve the presentation of the material. We sincerely thank Jun Liu for providing forecasts from FuXi-ENS. This work was supported by the National Natural Science Foundation of China, No. 42288101, 42375058, and the Academician Workstation Of AP-TCRC. The computations in this research were performed using the CFFF platform of Fudan University.

Author contributions

M.M., J.F., and J.P. designed the project. M.M. and J.F. managed and oversaw the project. J.F. and J.P. implemented and evaluated the perturbation generator. H.L. and X.Z. developed the FuXi model. J.P., J.F., and M.M. wrote and revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41612-025-01009-9>.

Correspondence and requests for materials should be addressed to Jie Feng.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025