

Revealing the Potential of Learnable Perturbation Ensemble Forecast Model for Tropical Cyclone Prediction

Jun Liu¹, Tao Zhou², Jiarui Li¹, Xiaohui Zhong¹, Peng Zhang², Jie Feng², Lei Chen^{1,3} and Hao Li^{1,3*}

¹Artificial Intelligence Innovation and Incubation Institute, Fudan University, Shanghai, 200433, China.

²Department of Atmospheric and Oceanic Sciences, Institute of Atmospheric Sciences, Fudan University, Shanghai, 200438, China.

³Shanghai Academy of Artificial Intelligence for Science, Shanghai, 200232, China.

*Corresponding author(s). E-mail(s): lihao_lh@fudan.edu.cn;

Contributing authors: junliu23@m.fudan.edu.cn;

tzhou22@m.fudan.edu.cn; jrli24@m.fudan.edu.cn;

x7zhong@gmail.com; zhang_peng@fudan.edu.cn;

fengjie@fudan.edu.cn; cltpys@163.com;

Abstract

Tropical cyclones (TCs) are highly destructive and inherently uncertain weather systems. Ensemble forecasting helps quantify these uncertainties, yet traditional systems are constrained by high computational costs and limited capability to fully represent atmospheric nonlinearity. FuXi-ENS introduces a learnable perturbation scheme for ensemble generation, representing a novel AI-based forecasting paradigm. Here, we systematically compare FuXi-ENS with ECMWF-ENS using all 90 global TCs in 2018, examining their performance in TC-related physical variables, track and intensity forecasts, and the associated dynamical and thermodynamical fields. FuXi-ENS demonstrates clear advantages in predicting TC-related physical variables, and achieves more accurate track forecasts with reduced ensemble spread, though it still underestimates intensity relative to observations. Further dynamical and thermodynamical analyses reveal that FuXi-ENS better captures

large-scale circulation, with moisture turbulent energy more tightly concentrated around the TC warm core, whereas ECMWF-ENS exhibits a more dispersed distribution. These findings highlight the potential of learnable perturbations to improve TC forecasting skill and provide valuable insights for advancing AI-based ensemble prediction of extreme weather events that have significant societal impacts.

Keywords: Tropical cyclone, Ensemble forecasting, Learnable perturbation, FuXi-ENS

1 Introduction

Tropical cyclones (TCs) are among the most destructive natural hazards on Earth [1], causing severe economic losses and casualties annually [2–6]. As highly nonlinear systems, TCs exhibit pronounced uncertainty [7]. While deterministic forecasts provide only single-trajectory predictions, ensemble forecasts quantify these uncertainties by sampling initial condition and model errors and accounting for day-to-day error growth variability [8–12]. Given the destructive impacts and intrinsic unpredictability of TCs, improving ensemble forecasting systems is critical for advancing TC prediction capabilities.

The design of the perturbation scheme is essential for ensemble forecasting to achieve uncertainty quantification and improve prediction skill [13, 14]. Traditional ensemble systems are built upon numerical weather prediction (NWP) models, where perturbation schemes are designed to represent uncertainties in initial conditions and physical parameterizations [13, 15]. The European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble prediction system (ECMWF-ENS) employs singular vectors (SV) to identify the fastest-growing perturbation modes for representing initial condition uncertainties [16, 17], and incorporates the Stochastically Perturbed Parametrization Tendencies (SPPT) scheme to characterize model physics uncertainties [18]. This design yields exceptional skill in mid-latitude 500 hPa geopotential forecasts, with high anomaly correlation coefficients (ACC) in Northern Hemisphere mid-latitudes [9, 19–21]. However, significant biases remain in TC forecasting [22, 23]. The National Centers for Environmental Prediction Global Ensemble Forecast System (NCEP-GEFS) generates initial perturbations using the ensemble Kalman filter (EnKF) approach [24, 25]. GEFS performs well in short-to-medium range TC track forecasts [26, 27], yet systematic biases persist in TC intensity prediction [28, 29]. Hybrid data assimilation methods that combine Four-Dimensional Variational (4D-Var) and EnKF techniques have improved TC initialization and ensemble forecasting by leveraging flow-dependent background error covariance while maintaining the dynamical constraints of variational frameworks [30, 31]. These hybrid systems exhibit enhanced performance in capturing complex TC structural evolution. Despite such progress [32], conventional ensemble forecasting remains computationally

expensive, with costs scaling linearly with ensemble size and imposing heavy burdens on high-performance computing resources [9, 33, 34], thereby limiting its scalability, operational feasibility, and further development.

Recent advances in machine learning have revolutionized weather forecasting, significantly alleviating the computational demands of NWP models while maintaining high accuracy. Models such as FourCastNet [35], Pangu-Weather [36], GraphCast [37], and FuXi [38] achieve deterministic forecasts with remarkable computational efficiency and skill. However, deterministic models inherently lack the ability to quantify forecast uncertainty, limiting their use in probabilistic forecasting [39]. To overcome this limitation, AI-based ensemble forecasting systems have emerged. Diffusion-based GenCast [40] generates ensemble forecasts through multiple sampling iterations and exhibits exceptional probabilistic forecasting skill. Yet its multi-step denoising processes necessitates high computational costs relative to other machine learning models [41]. Lang et al. [8] proposed AIFS-CRPS, which incorporates the Continuous Ranked Probability Score (CRPS) as the loss function for ensemble forecasting. To mitigate bias from finite ensemble sizes during training, they introduced the almost fair CRPS. However, the reliance on limited ensemble members may still introduce sampling limitations. FuXi-ENS [41] introduces a flow-dependent, learnable perturbation scheme that dynamically optimizes perturbations through data-driven learning [42, 43], offering a novel AI-based ensemble forecasting paradigm. Nevertheless, a systematic evaluation of this paradigm for TC prediction has not been conducted.

Here, we establish a comprehensive evaluation framework using a global dataset of 90 TCs from 2018 to systematically compare FuXi-ENS with ECMWF-ENS, the current state-of-the-art physics-based ensemble forecasting system. Our evaluation follows a progressive, multi-dimensional approach: (1) assessing forecast skill for key TC-related physical variables; (2) analyzing TC track prediction accuracy, including along-track and cross-track error evolution and the consistency between ensemble mean position errors and ensemble spread; (3) evaluating TC strike probability and intensity forecast skill; and (4) diagnosing the dynamical and thermodynamical mechanisms underlying FuXi-ENS performance through analyses of large-scale circulation evolution and perturbation energy distributions [44, 45].

2 Results

In this study, we evaluate FuXi-ENS using 90 TCs from 2018, using ERA5 reanalysis data as initial conditions and performing twice-daily forecasts at 00 UTC and 12 UTC. FuXi-ENS produces ensemble forecasts with 48 members, while ECMWF-ENS contains 51 members. Both systems provide 15-day forecasts at 0.25° spatial resolution with 6-hour intervals, covering all 90 global TCs in 2018. A comprehensive, multi-dimensional evaluation is conducted, including analyses of TC-related physical variables, ensemble track forecast errors, TC strike probability and intensity forecast skill, and the underlying dynamical

and thermodynamical fields. In this evaluation, ERA5 reanalysis data and the International Best Track Archive for Climate Stewardship (IBTrACS) records serve as ground truth for meteorological field and track validation, respectively.

2.1 TC-Related Physical Variables Performance

To comprehensively evaluate FuXi-ENS performance in TC prediction, we first assess its forecast skill for key physical variables closely related to TC behavior. We examine six variables spanning surface and upper-level processes: mean sea level pressure (MSL), 2m temperature (T2M), 10m wind speed (WS10M), 500 hPa geopotential (Z500), 850 hPa temperature (T850), and 850 hPa wind speed (WS850). These variables capture critical aspects of TC dynamics: MSL reflects the pressure gradient driving TC intensity; T2M and T850 characterize the thermal structure essential for TC development; WS10M and WS850 represent wind circulation patterns governing TC organization; and Z500 captures the large-scale flow patterns influencing TC motion [46–48]. To quantify probabilistic forecast skill, we employ the Receiver Operating Characteristic Area Skill Score (ROCASS), which measures a model’s ability to discriminate extreme events [49]. Scores are computed across all lead times (6–360 h) for all 730 initialization times in 2018 (00 and 12 UTC daily), separately for upper and lower terciles of each variable.

Figs. 1 and 2 present ROCASS for the upper and lower terciles of surface and upper-level variables, respectively. For both FuXi-ENS and ECMWF-ENS, ROCASS declines progressively with increasing forecast lead time, reflecting the expected deterioration of forecast skill across all variables examined. This degradation is steepest for wind fields (WS10M, WS850), whose ROCASS values fall below 0.2 by day 15 (Fig. 1e, f; Fig. 2e, f), indicating greater uncertainty and relatively lower predictability. In contrast, mean sea level pressure (MSL), temperature (T2M, T850), and geopotential fields (Z500) maintain higher skill, with ROCASS values remaining near 0.4 at day 15 (Fig. 1a–d; Fig. 2a–d). FuXi-ENS demonstrates marginally higher ROCASS values for wind speed variables (WS10M, WS850) compared to ECMWF-ENS, with the difference being more pronounced for both upper and lower extreme terciles, indicating superior ability to discriminate both strong and weak extreme wind events. For T2M tercile events, however, FuXi-ENS outperforms ECMWF-ENS through day 8 but exhibits slightly lower skill beyond day 9.

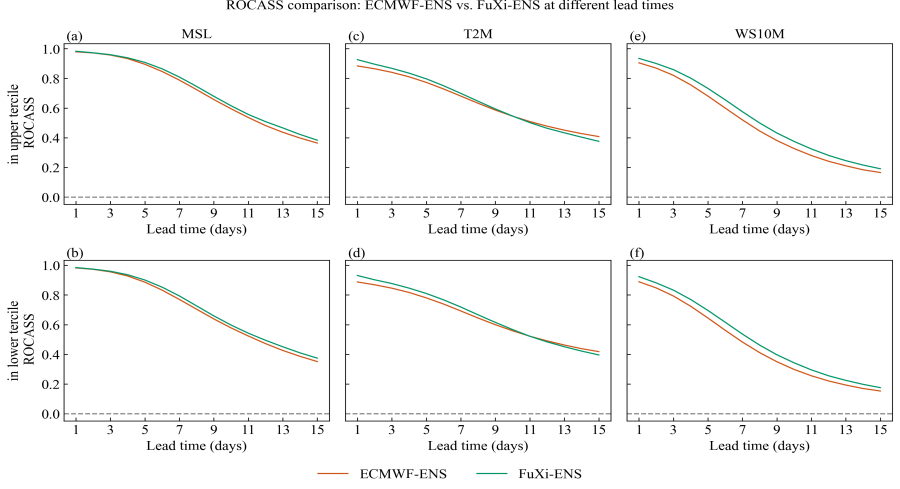


Fig. 1: ROCASS comparison between FuXi-ENS (green) and ECMWF-ENS (orange) for surface variables over 15-day forecasts with 6-hourly evaluation intervals. ROCASS are evaluated for the upper tercile (top panel) and lower tercile (bottom panel) of three surface variables: mean sea level pressure (MSL, left panel), 2m temperature (T2M, middle panel), and 10m wind speed (WS10M, right panel).

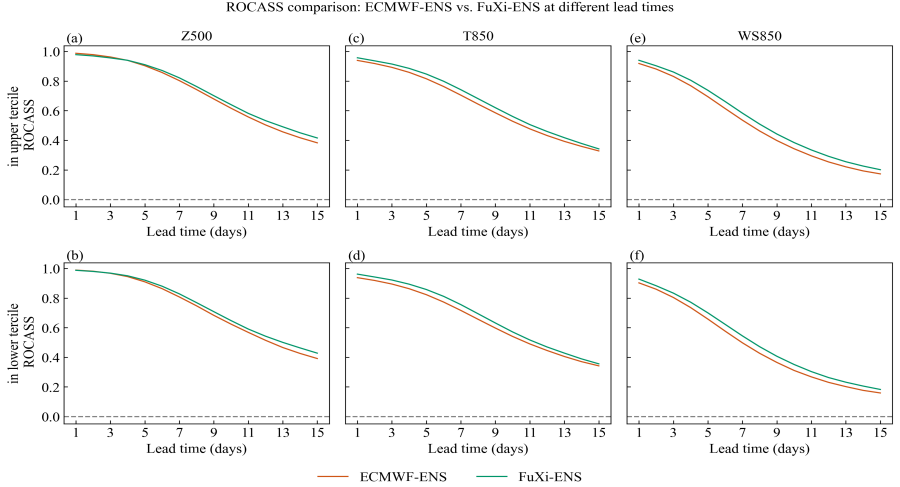


Fig. 2: ROCASS comparison between FuXi-ENS (green) and ECMWF-ENS (orange) for upper-level variables over 15-day forecasts with 6-hourly evaluation intervals. ROCASS are evaluated for the upper tercile (top panel) and the lower tercile (bottom panel) of three upper-level atmospheric variables: 500 hPa geopotential (Z500, left panel), 850 hPa temperature (T850, middle panel), and 850 hPa wind speed (WS850, right panel).

2.2 Ensemble Forecast Track Error Analysis for TCs

Building upon the comprehensive ROCASS evaluation that demonstrated FuXi-ENS’s superior performance in TC-related physical variables, we now present a detailed track error analysis for TC ensemble forecasts. Our TC tracking algorithm (Supplementary Fig. 10) identifies cyclone centers based on minimum MSL and maximum relative vorticity at 850 hPa. We apply this algorithm to all 90 TCs in 2018, tracking FuXi-ENS prediction fields from initialization throughout their life cycles (00 and 12 UTC daily), yielding track forecasts for each ensemble member. Accumulated ensemble mean position error ($\text{AccError}_{\text{TC}}$) quantifies the deviation between ensemble mean forecasts and IBTrACS observations, while accumulated ensemble spread ($\text{AccSpread}_{\text{TC}}$) measures the dispersion of individual ensemble members relative to the ensemble mean. To elucidate directional biases, we further decompose ensemble mean track errors into along-track error (AT, the component parallel to the observed track direction) and cross-track error (CT, the component perpendicular to the observed track direction), as illustrated in Supplementary Fig. 11.

Fig. 3 reveals the temporal evolution of four metrics across forecast lead times: $\text{AccError}_{\text{TC}}$, $\text{AccSpread}_{\text{TC}}$, AT, and CT. While both systems exhibit increasing $\text{AccError}_{\text{TC}}$ with lead time, FuXi-ENS consistently achieves significantly lower median errors and markedly tighter error distributions than ECMWF-ENS (Fig. 3a). $\text{AccSpread}_{\text{TC}}$ similarly increases with lead time, yet FuXi-ENS maintains considerably smaller spread than ECMWF-ENS throughout all forecast horizons (Fig. 3b). Directional error decomposition provides further insights into forecast biases: AT quantifies TC positional deviation parallel to the observed motion, with positive values indicating forecasts ahead of observations and negative values indicating lag. Both systems exhibit systematic lag bias (Fig. 3c), but ECMWF-ENS shows substantially greater median lag and broader error distributions. In contrast, CT quantifies deviation perpendicular to the observed motion, exhibits near-zero medians for both systems, indicating negligible systematic lateral bias, though ECMWF-ENS displays considerably wider distributions (Fig. 3d), reflecting greater forecast uncertainty in the lateral dimension.

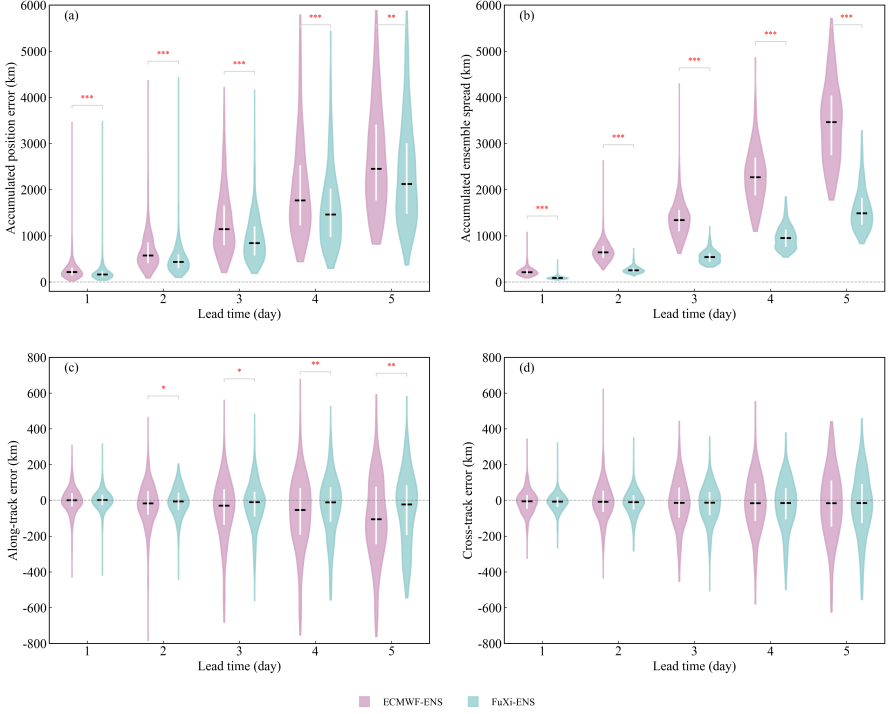


Fig. 3: Violin plots of TC track forecast errors in 5-day forecasts for ECMWF-ENS (purple) and FuXi-ENS (teal). (a), accumulated ensemble mean position error ($\text{AccError}_{\text{TC}}$, km). (b), accumulated ensemble spread ($\text{AccSpread}_{\text{TC}}$, km). (c), along-track error (AT, km). (d), cross-track error (CT, km). Violin plots show the distribution characteristics of each error metric across forecast lead times, with dashed horizontal lines indicating the median and interquartile ranges marked. Purple represents ECMWF-ENS forecasts, and teal represents FuXi-ENS forecasts. Asterisks indicate statistical significance levels (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$) based on Mann-Whitney U tests comparing the two ensemble systems.

Fig. 4 further compares the relationships between $\text{AccError}_{\text{TC}}$ and $\text{AccSpread}_{\text{TC}}$, as well as between AT and CT for both systems. FuXi-ENS exhibits more concentrated scatter points, with most samples showing $\text{AccError}_{\text{TC}}$ exceeding $\text{AccSpread}_{\text{TC}}$ (Fig. 4a). In contrast, ECMWF-ENS displays wider scatter distribution, with most samples showing $\text{AccSpread}_{\text{TC}}$ exceeding $\text{AccError}_{\text{TC}}$ (Fig. 4b). For the relationship between AT and CT, scatter points closer to the origin indicate smaller systematic bias. FuXi-ENS shows more concentrated distribution near the origin with relatively uniform scatter across four quadrants, indicating smaller biases in both directions (Fig. 4c). In contrast, ECMWF-ENS exhibits more dispersed scatter points with

significantly more samples in the negative AT direction, revealing obvious lagging bias. Additionally, ECMWF-ENS shows more outliers, further reflecting forecast instability (Fig. 4d). In summary, FuXi-ENS outperforms ECMWF-ENS in TC track simulation with smaller accumulated ensemble mean position errors, more concentrated ensemble forecasts, and more reasonable along-track error characteristics.

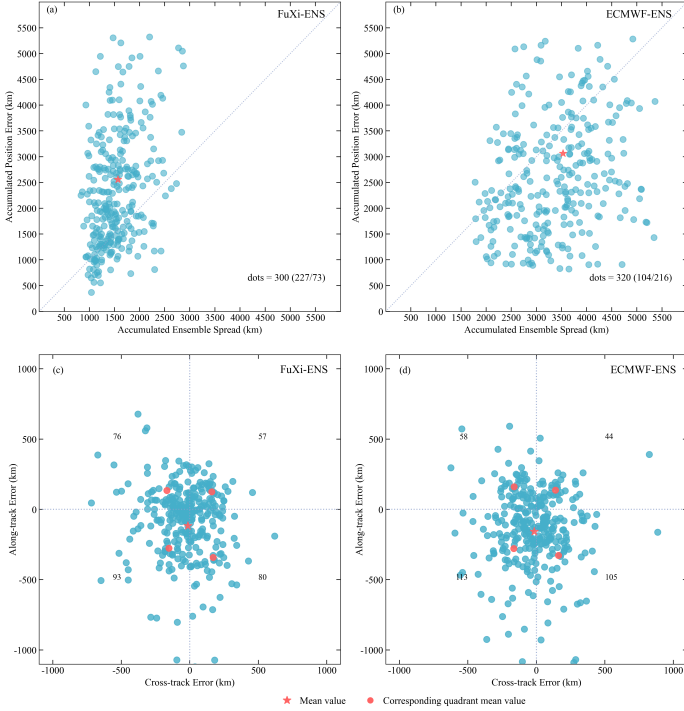


Fig. 4: Scatter plots comparing $\text{AccError}_{\text{TC}}$ against $\text{AccSpread}_{\text{TC}}$, and AT against CT at a 5-day forecast lead time for FuXi-ENS and ECMWF-ENS. Scatter plot distribution of $\text{AccError}_{\text{TC}}$ versus $\text{AccSpread}_{\text{TC}}$ for (a) FuXi-ENS and (b) ECMWF-ENS, and AT versus CT for (c) FuXi-ENS and (d) ECMWF-ENS. Each dot represents an individual TC forecast initialization. Pentagrams indicate overall means; in upper panels, dashed lines show 1:1 references with point counts for upper/lower regions; in lower panels, solid circles show quadrant means, numbers indicate counts per quadrant, and dashed lines show zero-error references.

2.3 TC strike probability and intensity forecast skill

Given FuXi-ENS's superior overall performance in TC-related physical variables and track error analysis, we further examine how these improvements manifest in specific individual TC cases. Taking typhoon KONG-REY as

an example, this typhoon formed in the central Northwest Pacific in late September, initially following a typical westward track, gradually moving northwestward and approaching east of the Philippines. From September 30 to October 2, it gradually turned northward; around October 4, it further shifted to a northeastward direction and finally approached the southern part of the Korean Peninsula on October 6 before dissipating. The evolution of this TC was influenced by the complex interaction between the subtropical high and mid-latitude trough systems, resulting in complex track changes but with typical representativeness.

The ensemble forecast results demonstrate that FuXi-ENS closely reproduces the best track (IBTrACS), accurately capturing the complete evolution sequence of westward movement, northward recurvature, and northeastward turning, particularly during the critical turning phase (October 2–4). Notably, the ensemble probability distribution remains relatively concentrated (Fig. 5a). In contrast, ECMWF-ENS exhibits larger simulation deviations, with its ensemble mean track showing a pronounced westward bias during the initial phase. Substantial deviations also occur during the recurvature phase, failing to accurately represent the typhoon’s northeastward turning position and recurvature magnitude. Furthermore, ECMWF-ENS displays a significantly wider ensemble probability distribution than FuXi-ENS, with particularly strong track dispersion following northward recurvature (Fig. 5b). This comparative analysis of Typhoon KONG-REY demonstrates that FuXi-ENS significantly outperforms ECMWF-ENS in both track simulation accuracy and ensemble probability distribution concentration. This superior performance is further corroborated by comparative analyses of representative cases, including Hurricane Florence (Atlantic Basin) and Hurricane Bud (Eastern Pacific Basin) (Supplementary Figs. 1, 2), highlighting the model’s robust generalization capability across different oceanic basins.

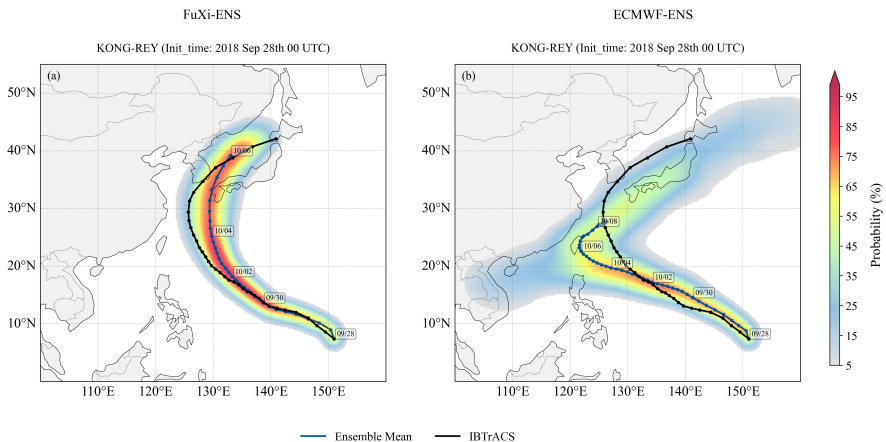


Fig. 5: TC strike probability forecast comparison for Typhoon KONG-REY. Probabilistic distribution of TC strike probability for Typhoon KONG-REY based on ensemble forecasts initialized at 00 UTC on September 28, 2018 from (a) FuXi-ENS and (b) ECMWF-ENS. Blue lines denote ensemble mean tracks; black lines indicate best track observations (IBTrACS). Shading indicates TC strike probability (%). Date labels mark the temporal evolution of the track.

Beyond TC track forecasting and strike probability prediction, accurate intensity prediction is equally critical. Statistical verification across all cases reveals that FuXi-ENS systematically underestimates TC intensity relative to both ECMWF-ENS and best track observations (IBTrACS), reflecting a persistent challenge in current AI-based weather prediction models [41, 50, 51]. To systematically evaluate the ensemble forecasting capability of FuXi-ENS for TCs, we present a representative case examining intensity evolution. ERA5 reanalysis data are used as the verification benchmark to ensure consistency with subsequent circulation analyses. Fig. 6 illustrates the intensity evolution of Typhoon KONG-REY, with corresponding results for Typhoon TRAMI presented in Supplementary Fig. 3. For the 10m maximum wind speed evolution, FuXi-ENS closely tracks ERA5 throughout the development phase (days 0–4) and accurately captures the timing of peak intensity during days 4–5, while maintaining a relatively narrow ensemble spread. In contrast, ECMWF-ENS exhibits considerably greater ensemble dispersion (Fig. 6a, b). The minimum central pressure evolution displays similar characteristics. FuXi-ENS reproduces the pressure deepening process closely matching ERA5 during days 0–4 and accurately captures the timing of minimum pressure, while maintaining a concentrated ensemble distribution. Conversely, ECMWF-ENS shows substantially greater ensemble dispersion in pressure evolution, indicating higher inter-member uncertainty (Fig. 6c, d).

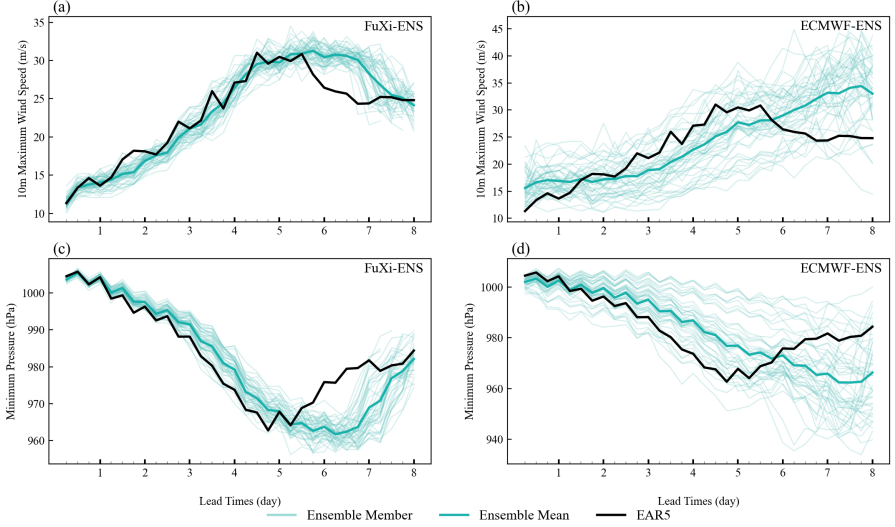


Fig. 6: Intensity forecast comparison for typhoon KONG-REY. 10m maximum wind speed variation with forecast lead time for (a) FuXi-ENS and (b) ECMWF-ENS initialized at 00 UTC on September 28, 2018. Minimum pressure variation with forecast lead time for (c) FuXi-ENS and (d) ECMWF-ENS. Light green thin lines represent ensemble member forecasts, dark green thick lines represent ensemble mean forecasts, and black thick lines represent ERA5 reanalysis forecasts.

2.4 TC-Related Dynamical Analysis

Based on the preceding analyses, FuXi-ENS outperforms ECMWF-ENS across both statistical metrics and case studies, more accurately reproducing the evolution of TC tracks and intensities. Since TC development is governed by complex dynamical and thermodynamical interactions, where large-scale steering flows primarily determine TC trajectories and sufficient moisture supports maintenance and intensification, we further examine whether FuXi-ENS more realistically simulates the associated atmospheric circulation and moisture distributions. This analysis elucidates the physical mechanisms underlying its superior performance.

Using KONG-REY as an example, FuXi-ENS more faithfully reproduces its “westward–northward–northeastward” track pattern compared to ECMWF-ENS. To diagnose the atmospheric circulation features responsible for these differences, we analyze the evolution of wind and geopotential height fields at multiple levels. Fig. 7 shows the 500 hPa wind speed (shading), geopotential height (contours), and horizontal wind vectors for KONG-REY initialized at

00 UTC on September 28, 2018, at lead times of 1, 3, 5, and 7 days from FuXi-ENS, ECMWF-ENS, and ERA5. Corresponding 850 hPa and surface level results are presented in Supplementary Figs. 4, 5. The 850 hPa circulation evolution for TRAMI case exhibits similar patterns (Supplementary Fig. 6).

At 500 hPa, the Western Pacific Subtropical High (WPSH) is the dominant system governing TC motion. Before October 5, both FuXi-ENS and ECMWF-ENS capture a reasonably strong and well-positioned WPSH ridge that steers KONG-REY northwestward, though FuXi-ENS more accurately represents the WPSH’s spatial extent compared to ERA5 (Fig. 7a-c, e-g, i-k). On October 5, both ERA5 reanalysis and FuXi-ENS show a deep trough over Japan and Korea, forming a “high–low cooperation” pattern with the WPSH that creates favorable steering flow for KONG-REY’s northward and subsequent eastward recurvature (Fig. 7d, l). In contrast, ECMWF-ENS fails to reproduce this trough, resulting in a relatively zonal circulation pattern that lacks the necessary recurvature mechanism (Fig. 7h).

Similar behavior is observed at 850 hPa. Prior to October 5, both systems reasonably simulate the WPSH’s position and the associated northwestward steering flow for KONG-REY. By October 5, FuXi-ENS maintains the curvature and position of the 1520 gpm geopotential height contour consistent with ERA5, whereas ECMWF-ENS produces a relatively zonal contour that inhibits the TC’s northeastward recurvature (Supplementary Fig. 4). At the surface, both ERA5 and FuXi-ENS exhibit pronounced curvature in the 1016 hPa isobar, reflecting the WPSH’s modulation of TC motion. In contrast, ECMWF-ENS displays relatively straight isobars, inadequately capturing the WPSH’s spatial structure and underestimating its intensity (Supplementary Fig. 5).

Likewise, FuXi-ENS also more accurately captures the large-scale circulation associated with Typhoon TRAMI (Supplementary Fig. 6). Overall, the dynamical analyses demonstrate that FuXi-ENS reproduces both TC tracks and the associated steering flows with higher fidelity, whereas ECMWF-ENS exhibits systematic biases in simulating key circulation systems, which may lead to inferior track forecasts.

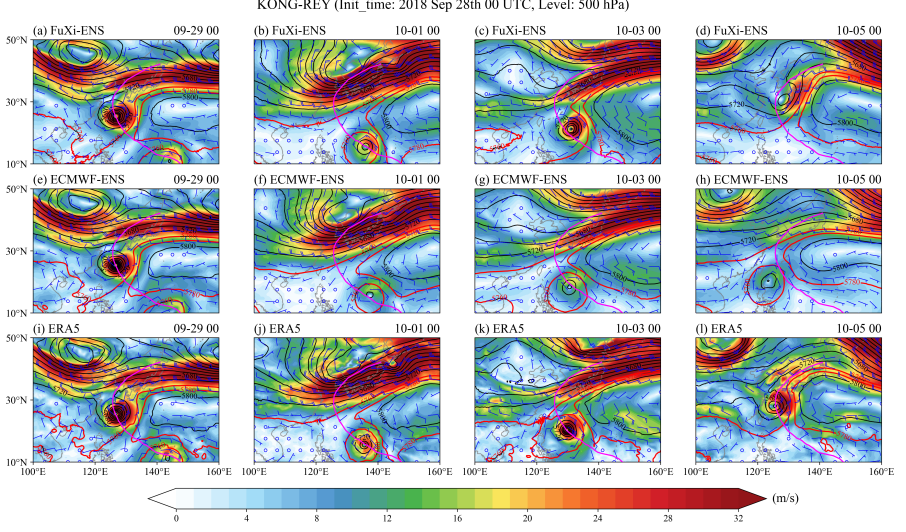


Fig. 7: Evolution of 500 hPa atmospheric circulation patterns for typhoon KONG-REY with forecast lead time. Evolution of KONG-REY’s 500 hPa wind speed (shading, m/s), geopotential height (contours, gpm, 40 gpm intervals), and horizontal wind (vectors, m/s) initialized at 00 UTC on September 28, 2018, for September 29 (first column), October 1 (second column), October 3 (third column), and October 5 (fourth column) from FuXi-ENS (first row), ECMWF-ENS (second row), and ERA5 (third row). Red contours indicate the 5760 and 5780 gpm contour line, and magenta lines represent best tracks (IBTrACS).

We further compare the ensemble mean and individual member simulations of specific geopotential height contour evolution for KONG-REY with consistent initialization time and lead times, as shown in Fig. 8. Similar results for TRAMI are shown in Supplementary Fig. 7.

Although both ensembles show increasing spread with lead time, FuXi-ENS members remain more tightly clustered, indicating greater internal consistency (Fig. 8a–d). In contrast, ECMWF-ENS exhibits substantially greater dispersion, with some members showing considerable deviations that reduce the representativeness of the ensemble mean (Fig. 8e–h). These results demonstrate that FuXi-ENS not only achieves higher accuracy but also maintains superior coherence among ensemble members.

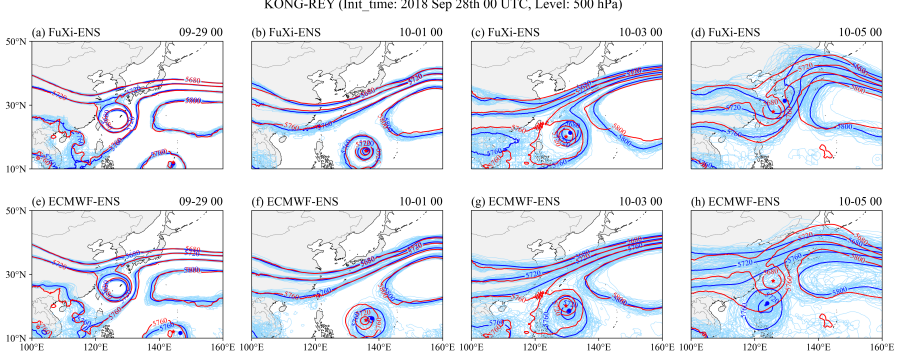


Fig. 8: Evolution of 500 hPa geopotential height ensemble forecasts for typhoon KONG-REY with forecast lead time. Evolution of KONG-REY’s 500 hPa geopotential height (contours, gpm) initialized at 00 UTC on September 28, 2018, for September 29 (first column), October 1 (second column), October 3 (third column), and October 5 (fourth column) from FuXi-ENS (first row), ECMWF-ENS (second row). Light blue lines represent individual ensemble members, dark blue lines represent the ensemble mean, and red lines represent ERA5 reanalysis. Four contour lines are shown: 5680 gpm, 5720 gpm, 5760 gpm and 5800 gpm. Blue dots denote the ensemble mean track and red stars denote the best track (IBTrACS).

2.5 TC-Related Thermodynamical Analysis

The evolution of TCs depends not only on accurate representation of dynamical fields but also critically on the evolution of thermal fields. Moisture turbulent energy (MTE), a key diagnostic of atmospheric thermal and moisture states, effectively reflects the energy distribution related to TCs.

Fig. 9 shows the evolution of 850 hPa MTE (shading), temperature (contours), and horizontal wind (vectors) for KONG-REY from FuXi-ENS and ECMWF-ENS with the same initialization time and lead times. Initially, at lead times of 1–3 days (September 29–October 1), both FuXi-ENS and ECMWF-ENS capture enhanced MTE near the TC warm core. However, ECMWF-ENS also produces extensive high MTE values across the mid- to high-latitude regions, while FuXi-ENS exhibits more compact MTE concentrated around the TC core with weaker peripheral perturbations (Fig. 9a, b, e, f). As the forecast lead time extends, the TC moves northward and begins to interact with mid-latitude frontal zones and jet streams. During this transition, at a lead time of 5 days (October 3), ECMWF-ENS generates stronger and more continuous high-MTE bands, indicating greater sensitivity to frontal perturbations and higher uncertainty, whereas FuXi-ENS maintains weaker and more localized MTE (Fig. 9c, g). In the later forecast period, at a lead time of 7 days (October 5), as the TC structure gradually dissipates, FuXi-ENS

exhibits significantly enhanced MTE with perturbation energy mainly concentrated in the mid-high latitude jet region, while ECMWF-ENS continues to maintain high MTE across both frontal and jet regions (Fig. 9d, h). Similar patterns are observed for FLORENCE and TRAMI (Supplementary Figs. 8 and 9). Overall, FuXi-ENS concentrates MTE evolution within the TC core, while ECMWF-ENS shows broader perturbation diffusion across larger spatial scales, which may contribute to its larger ensemble spread. These findings provide valuable guidance for refining ensemble perturbation schemes.

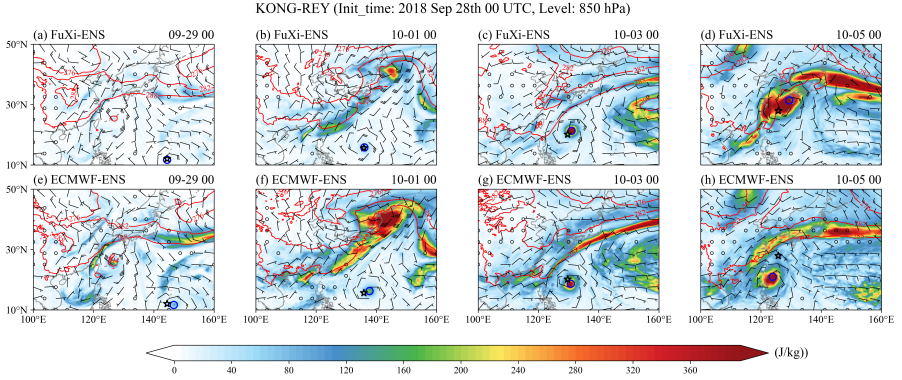


Fig. 9: Evolution of 850 hPa MTE for typhoon KONG-REY with forecast lead time. Evolution of KONG-REY’s 850 hPa MTE (shading, J/kg), temperature (contours, K), and horizontal wind (vectors, m/s) initialized at 00 UTC on September 28, 2018, for September 29 (first column), October 1 (second column), October 3 (third column), and October 5 (fourth column) from FuXi-ENS (first row), ECMWF-ENS (second row). Blue circles denote the ensemble mean track and black stars denote the best track (IBTrACS).

3 Discussion

Systematic evaluation of FuXi-ENS’s learnable perturbation generation scheme demonstrates significant advantages across multiple dimensions of TC forecasting. For TC-related physical variables, FuXi-ENS generally achieves superior performance relative to ECMWF-ENS. Track forecast analysis reveals that FuXi-ENS exhibits substantially lower accumulated ensemble mean position errors and reduced ensemble spread compared to ECMWF-ENS. For intensity forecasts, FuXi-ENS reproduces TC intensity evolution more faithfully than ECMWF-ENS when evaluated against ERA5 reanalysis data. Dynamic and thermodynamic analyses further elucidate the physical mechanisms underlying FuXi-ENS’s superior performance. Using Typhoon KONG-REY as a case study, FuXi-ENS more accurately captures the evolution of

steering flows associated with the TC, including interactions between the subtropical high and mid-latitude troughs. Additionally, detailed analysis of MTE distribution reveals distinct characteristics between the two ensemble systems: FuXi-ENS exhibits highly concentrated MTE within the TC warm-core region, whereas ECMWF-ENS demonstrates broader spatial MTE distribution. These differences may directly explain the smaller ensemble spread in FuXi-ENS, providing valuable insights for optimizing ensemble perturbation schemes in future operational systems.

The superior track forecasts but relatively weaker intensity predictions of FuXi-ENS reflect a common trade-off in current machine learning-based forecasting systems. The improved track accuracy is primarily attributed to FuXi-ENS’s ability to capture large-scale steering flow patterns through flow-dependent learnable perturbations, with this advantage being most pronounced at short-to-medium forecast lead times. In contrast, the limited skill in intensity prediction can be attributed to two fundamental challenges. First, systematic underestimation of TC intensities in ERA5 relative to observational benchmarks introduces biased training targets that fundamentally limit model capability. Second, the prevalent use of L1 or L2 loss functions introduces a “double penalty” problem [52], whereby spatially smoothed predictions are systematically favored over sharp intensity gradients. Future improvements may benefit from three key developments: (1) loss functions that mitigate the double penalty problem while preserving sensitivity to intensity gradients; (2) incorporation of higher-resolution datasets to better resolve extreme events; and (3) end-to-end optimization strategies specifically tailored for extreme event prediction. This approach would require leveraging high-quality, high-resolution datasets (such as high-resolution extreme event datasets) as training targets, with optimization strategies tailored to specific physical processes. Such targeted training may enable more precise characterization of atmospheric circulation patterns and improved capture of fine-scale features critical to extreme events. This paradigm shift may ensure that models better understand and accurately predict extreme weather phenomena that have significant societal impacts.

Overall, this study demonstrates that machine learning-based learnable perturbation techniques represent a significant advancement in TC ensemble forecasting. Beyond the demonstrated improvements in track prediction accuracy, FuXi-ENS exhibits enhanced capability in representing the complex dynamical and thermodynamical structures of TCs. These findings suggest that learnable perturbation schemes offer a promising pathway toward more reliable operational forecasting systems, with potential applications to high-impact extreme weather events that require ensemble uncertainty quantification.

4 Methods

4.1 Data

This study establishes a comprehensive evaluation framework to systematically assess the performance of the FuXi-ENS learnable perturbation-based ensemble forecasting system for tropical cyclone prediction. The framework integrates initial condition perturbations, benchmark model comparisons, and observational verification, providing a scientific foundation for objectively evaluating flow-dependent perturbation schemes.

As a deep learning-based ensemble system, FuXi-ENS depends critically on high-quality initial conditions. Here, we adopt the fifth generation of the European Centre for Medium-Range Weather Forecasts (ECMWF) reanalysis dataset (ERA5) [53] as model input. ERA5 provides atmospheric fields at 0.25° spatial resolution with 6-hourly temporal resolution, delivering accurate atmospheric states for machine learning applications. Initialized with ERA5 fields, FuXi-ENS generates 15-day ensemble forecasts at 6-hour intervals.

For benchmarking, we select the ECMWF ensemble prediction system (ECMWF-ENS) as a reference. ECMWF-ENS has been extensively validated in previous studies and is widely regarded as one of the most authoritative operational ensemble forecasting systems for tropical cyclone track prediction. Historical ECMWF-ENS forecasts are obtained from the Weather-Bench2 platform ([gs://weatherbench2/datasets/ifs_ens/2018-2022-1440x721.zarr](https://weatherbench2/datasets/ifs_ens/2018-2022-1440x721.zarr)), covering the period 2018–2022 at 1440×721 spatial resolution with 6-hourly output intervals.

Tropical cyclone track and intensity predictions serve as core components of forecast evaluation. For FuXi-ENS, a tracking algorithm is employed to extract TC tracks and intensities, with detailed descriptions provided in Section 4.3. Both ECMWF high-resolution deterministic forecasts (ECMWF-HRES) and ensemble forecasts (ECMWF-ENS) TC track data are retrieved from the TIGGE (THORPEX Interactive Grand Global Ensemble) archive in XML format (<https://confluence.ecmwf.int/display/TIGGE/Tools>).

For observational verification, we use the International Best Track Archive for Climate Stewardship (IBTrACS) as ground truth (<https://www.ncei.noaa.gov/products/international-best-track-archive>). Maintained by the World Meteorological Organization, IBTrACS compiles official best track records of global tropical cyclones, including 3-hourly center locations, maximum sustained winds, and minimum central pressures [54, 55]. The dataset undergoes rigorous quality control and multi-agency cross-validation, establishing it as the authoritative reference for TC research. Comparative analysis against IBTrACS enables objective assessment of track accuracy, intensity prediction skill, and probabilistic forecast performance, providing robust evidence for evaluating learnable perturbation techniques in extreme weather forecasting.

4.2 FuXi-ENS model

This study employs the FuXi-ENS model to explore its performance in TC track prediction. The overall framework and evaluation methodology are illustrated in Fig. 10. FuXi-ENS leverages machine learning to address ensemble forecasting challenges, offering fast and efficient probabilistic predictions through learnable perturbations that capture flow-dependent characteristics. Previous studies have demonstrated its superior performance relative to ECMWF-ENS across multiple probabilistic metrics, including continuous ranked probability score (CRPS) and Brier score [41]. In this work, we comprehensively assess FuXi-ENS's forecasting skill for tropical cyclones using a dataset of 90 TCs from 2018. The technical architecture and key innovations of FuXi-ENS are described below.

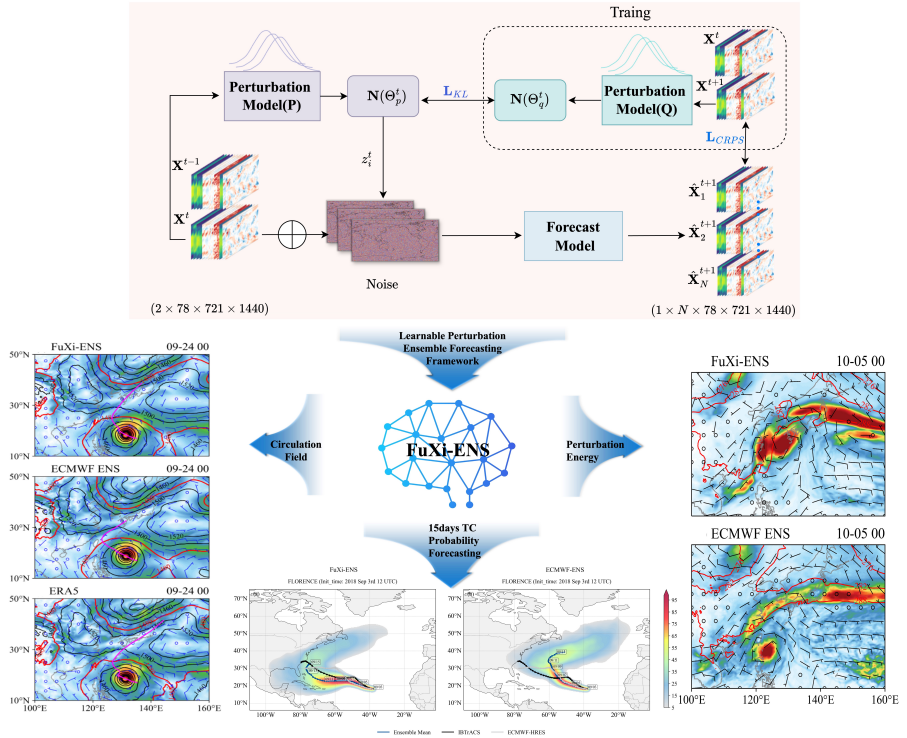


Fig. 10: Schematic Diagram of Learnable Perturbation Ensemble Forecasting Framework and TC Prediction Performance Evaluation. The framework includes a perturbation generation module (upper panel) and a forecast model, with evaluation based on atmospheric circulation analysis (left) and perturbation energy assessment (right). Performance is compared between FuXi-ENS, ECMWF-ENS, and ERA5, demonstrating the effectiveness of learnable perturbation methods in TC ensemble prediction.

FuXi-ENS adopts a dual-module design architecture. The perturbation generation module is based on a variational autoencoder (VAE) framework [56, 57]. It encodes multi-dimensional meteorological fields into multivariate Gaussian probability distributions. Perturbation vectors are generated through random sampling from these distributions and superimposed onto initial atmospheric states to form perturbed initial fields. The forecasting module employs an encoder-decoder structure that receives perturbed initial fields as input. Combined with Swin Transformer architecture [58], it generates future forecasts through autoregressive processes. The model’s most significant technical innovation lies in its flow-dependent perturbation strategy. It introduces perturbations not only at initial time but regenerates flow-dependent perturbations at each forecast time step, superimposing them onto current states. This achieves a dynamic perturbation update mechanism. The design enables the model to adaptively adjust perturbation patterns based on real-time atmospheric evolution characteristics, more accurately capturing uncertainty propagation in nonlinear systems like TCs. FuXi-ENS is trained on ERA5 reanalysis data (2002-2016). Through multiple random sampling and autoregressive forecasting processes, it generates 15-day ensemble forecasts with 48 ensemble members at 0.25° spatial resolution and 6-hour temporal resolution. The model covers 5 upper-air variables (geopotential (Z), temperature (T), u component of wind (U), v component of wind (V), specific humidity (Q)) and 13 surface variables (2m temperature (T2M), 10m u wind component (U10M), 10m v wind component (V10M), mean sea level pressure (MSL), total precipitation (TP), etc.), providing crucial technical foundation for probabilistic assessment of TC track and intensity prediction. Further details on the FuXi-ENS model can be found in Zhong et al. [41].

4.3 TC tracking method

The method employed in this study is adapted from the ECMWF TC tracking algorithm developed by Zhong et al. [59, 60]. The logic of the tracking algorithm is illustrated in Supplementary Fig. 10. First, a preliminary estimate of the TC position is obtained based on observational data. Within a 445 km radius surrounding this estimated location, candidate points are generated by identifying local minima in MSL and local maxima in 10-meter wind vorticity. Among these candidates, the algorithm selects the nearest point that satisfies a set of predefined criteria (detailed in Supplementary Table 1) as the identified TC position. If none of the candidate points meet the criteria, the TC is considered to be absent and the procedure terminates. Subsequently, a potential TC position for the next forecast time step is estimated based on the updated position. This estimation combines two approaches: linear extrapolation using the TC positions from the two previous time steps, and the computation of an advection vector derived from wind fields at a specific pressure level. Thereafter, The algorithm applies the same procedure at each time step to identify TC positions.

At the first forecast time step, where only a single historical position is available, the TC position is estimated solely using the advection vector scheme. To mitigate the influence of false local minima caused by the inherent smoothness of machine learning-based forecasts, the ensemble forecast is first downsampled from 0.25° to 1.25° before candidate generation and then upsampled back to its original resolution. Furthermore, to avoid excessively large displacements, the TC position change is constrained to no more than three times the displacement in the previous time step.

4.4 Evaluation metrics

ROCASS: The Receiver Operating Characteristic (ROC) curve is a widely used metric for evaluating the discriminative ability of binary event forecasts, particularly in probabilistic ensemble prediction [61]. By varying the decision threshold, it characterizes the trade-off between the probability of detection (POD) and the probability of false detection (POFD), thereby quantifying the model’s ability to distinguish between “event” and “non-event” cases. Plotting POFD on the x-axis and POD on the y-axis yields the ROC curve. The area under the ROC curve (ROCA) provides a numerical summary of forecast skill, ranging from 0 to 1, where values closer to 1 indicate stronger discrimination, and 0.5 denotes no skill beyond random guessing. To facilitate comparison against a no-skill baseline, the ROCA is generally converted into a skill score, referred to as the ROCASS skill score (ROCASS) [49], defined as below:

$$\text{ROCASS} = 2 \times (\text{ROCA} - 0.5) \quad (1)$$

This transformation normalizes the score to the range $[-1, 1]$, where $\text{ROCASS} = 1$ denotes a perfect forecast, $\text{ROCASS} = 0$ indicates equivalence to a no-skill forecast, and $\text{ROCASS} < 0$ implies performance worse than random.

Accumulated Ensemble Mean Position Error and Accumulated Ensemble Spread: In evaluating TC ensemble track forecasts, it is essential to assess both forecast accuracy and the model’s ability to represent forecast uncertainty. Two primary metrics are commonly employed: ensemble mean position error (ErrorTC) and ensemble spread (SpreadTC). The ErrorTC, calculated as the root-mean-square error (RMSE) between the ensemble mean track and the observed best track (IBTrACS) [62], quantifies the deterministic accuracy of the ensemble mean forecast. The SpreadTC, computed as the RMSE between each ensemble member’s track and the ensemble mean track, measures the dispersion among members’ forecast trajectories. Ideally, SpreadTC should be consistent with the magnitude of ErrorTC [63], indicating that the model’s uncertainty estimation is reliable. The formulas are as follows:

$$\text{Error}_{\text{TC}}(t) = \sqrt{\frac{1}{N} \sum_{n=1}^N \left\| \bar{\mathbf{p}}_{\text{forecast}}^{(n,t)} - \mathbf{p}_{\text{o}}^{(n,t)} \right\|^2} \quad (2)$$

$$\text{Spread}_{\text{TC}}(t) = \sqrt{\frac{1}{N} \sum_{n=1}^N \frac{1}{M} \sum_{m=1}^M \left\| \mathbf{p}_{\text{forecast}}^{(n,m,t)} - \bar{\mathbf{p}}_{\text{forecast}}^{(n,t)} \right\|^2} \quad (3)$$

where N represents the number of samples, M denotes the number of ensemble members, $\bar{\mathbf{p}}_{\text{forecast}}^{(n,t)}$ represents the ensemble mean position for the n^{th} sample at forecast lead time t , $\mathbf{p}_o^{(n,t)}$ represents the observed best track position for the n^{th} sample at lead time t , and $\mathbf{p}_{\text{forecast}}^{(n,m,t)}$ represents the forecast position for the n^{th} sample, m^{th} member at lead time t .

By accumulating Error_{TC} and $\text{Spread}_{\text{TC}}$ across multiple forecast lead times, the Accumulated Ensemble Mean Position Error ($\text{AccError}_{\text{TC}}$) and Accumulated Ensemble Spread ($\text{AccSpread}_{\text{TC}}$) are obtained. These accumulated metrics are used to comprehensively evaluate the track prediction accuracy and overall uncertainty throughout the forecast period:

$$\text{AccError}_{\text{TC}} = \sum_{t \in T} \text{Error}_{\text{TC}}(t) \quad (4)$$

$$\text{AccSpread}_{\text{TC}} = \sum_{t \in T} \text{Spread}_{\text{TC}}(t) \quad (5)$$

Ideally, $\text{AccSpread}_{\text{TC}}$ should match $\text{AccError}_{\text{TC}}$. If $\text{AccSpread}_{\text{TC}} > \text{AccError}_{\text{TC}}$, it indicates an overestimation of forecast uncertainty. If $\text{AccSpread}_{\text{TC}} < \text{AccError}_{\text{TC}}$, it signifies an underestimation of forecast uncertainty.

Along-track Error and Cross-track Error: To quantify TC track forecast errors, this study decomposes Error_{TC} into Along-track error (AT) and Cross-track error (CT) [64, 65]. The calculation proceeds as follows (The detailed schematic diagram of the calculation principle is shown in Supplementary Fig. 11). Let two consecutive observed positions be denoted as OB_1 and OB_2 , and the forecast position as FC . AT represents the displacement of the forecast point in the tangential direction of the observed track and is defined as [66, 67]:

$$\text{AT} = \frac{\overrightarrow{\text{OB}_1\text{OB}_2} \cdot \overrightarrow{\text{OB}_1\text{FC}}}{\|\overrightarrow{\text{OB}_1\text{OB}_2}\|} - \|\overrightarrow{\text{OB}_1\text{OB}_2}\| \quad (6)$$

where $\overrightarrow{\text{OB}_1\text{OB}_2}$ is the track tangent vector, and $\overrightarrow{\text{OB}_1\text{FC}}$ is the vector from OB_1 to the forecast point. AT is defined as the projection error of the forecast position in the direction of the observed track: $\text{AT} > 0$ indicates the forecast position is ahead of the observed track, $\text{AT} < 0$ indicates the forecast position lags behind the observed track.

CT measures the displacement perpendicular to the observed track and is computed using the Pythagorean relation:

$$\text{CT} = \sqrt{\|\overrightarrow{\text{OB}_1\text{FC}}\|^2 - \left(\frac{\overrightarrow{\text{OB}_1\text{OB}_2} \cdot \overrightarrow{\text{OB}_1\text{FC}}}{\|\overrightarrow{\text{OB}_1\text{OB}_2}\|} \right)^2} \quad (7)$$

CT is defined as the deviation of the forecast position perpendicular to the observed track direction: $CT > 0$ indicates that the forecast position lies to the right of the observed track, while $CT < 0$ indicates that it lies to the left. Together, AT and CT provide a clear decomposition of forecast error into tangential and normal components, forming a quantitative basis for evaluating TC track prediction performance.

Mann-Whitney U test: To evaluate the statistical significance of performance differences between the two ensemble forecasting systems in TC prediction, we employed the Mann-Whitney U test. This non-parametric test is particularly well-suited for comparing TC forecast skill metrics between the ECMWF-ENS and FuXi-ENS systems, as it does not require the assumption of normality—a condition frequently violated by meteorological forecast verification scores. The Mann-Whitney U test assesses whether two independent samples originate from the same underlying distribution by comparing their rank-ordered values, thereby providing robustness against outliers that may occur in TC forecast performance data [68, 69].

Strike Probability: To quantify the uncertainty in TC forecasts, we calculate the strike probability using ensemble predictions. The computational method can be expressed as follows [70, 71]:

The strike probability for grid point (i,j) is calculated as:

$$P(i,j) = \frac{1}{N} \times \sum_{k=1}^N I_k(i,j) \times 100\% \quad (8)$$

where N is the total ensemble members. the impact zone indicator function I_k is defined as:

$$I_k(i,j) = \begin{cases} 1, & \text{if } d(\text{point}(i,j), \text{track}_k) \leq R \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

where d is the minimum distance from point to track; R is the impact radius (1° , approximately 111 km).

For multiple concurrent TCs, a maximum value merging strategy is applied:

$$P_{\text{total}}(i,j) = \max\{P_1(i,j), P_2(i,j), \dots, P_m(i,j)\} \quad (10)$$

where m is the number of concurrent TCs.

Moisture Turbulent Energy: To quantitatively assess the energy characteristics of perturbations in ensemble forecasts, we adopt a thermodynamic perspective of TC processes and employ Moisture Turbulent Energy (MTE) perturbations as a diagnostic metric. MTE comprehensively incorporates perturbation components of wind speed, temperature, and specific humidity, effectively reflecting the perturbation characteristics of atmospheric thermodynamic states within TC systems. The calculation formula is [72]:

$$\text{MTE}_{\text{TC}} = \frac{1}{2} \left[u'^2_{i,j,k} + v'^2_{i,j,k} \right] + \frac{c_p}{T_r} T'^2_{i,j,k} + \varepsilon \frac{L^2}{c_p T_r} q'^2_{i,j,k} \quad (11)$$

where u' and v' denote the zonal and meridional wind perturbation respectively, T' represents the temperature perturbation, and q' is the specific humidity perturbation. Each perturbation is defined as the deviation of an ensemble member forecast from the ensemble mean. The constants are specified as follows: reference temperature $T_r = 270$ K, specific heat of dry air at constant pressure $c_p = 1005.7 \text{ J kg}^{-1} \text{ K}^{-1}$, and latent heat of water vapor condensation $L = 2.51 \times 10^6 \text{ J kg}^{-1}$.

In this formulation, the first term represents the contribution from perturbation kinetic energy, the second term reflects the effect of temperature perturbations on available potential energy, and the third term captures the enhancement of perturbation energy through latent heat release associated with specific humidity. Compared with traditional dry-air perturbation energy measures, MTE provides a more comprehensive description of the role of moist processes in the growth of ensemble perturbations. In practice, MTE is evaluated at each horizontal grid point (i,j) and vertical level k , thereby yielding a three-dimensional distribution of perturbation energy.

Data availability

We downloaded a subset of the ERA5 dataset from the official Copernicus Climate Data Store (CDS) at <https://cds.climate.copernicus.eu/>. Additionally, ground truth TC tracks were obtained from the International Best Track Archive for Climate Stewardship (IBTrACS) project, publicly available at <https://www.ncei.noaa.gov/products/international-best-track-archive>. TC track forecasts from ECMWF-HRES (deterministic) and ECMWF-ENS (ensemble) were retrieved from the TIGGE (THORPEX Interactive Grand Global Ensemble) data archive at <https://confluence.ecmwf.int/display/TIGGE/Tools>. Furthermore, the historical operational ensemble forecast data used in this study were accessed from the WeatherBench2 platform's Google Cloud storage service at gs://weatherbench2/datasets/ifs_ens/2018-2022-1440x721.zarr.

Code availability

The FuXi-ENS model and related example code used for evaluation in this study are publicly available and can be accessed from Zenodo: <https://zenodo.org/records/15124541/files/FuXi-ENS-main.zip?download=1>.

Acknowledgements

We are grateful to ECMWF for providing the ERA5 dataset and acknowledge the ensemble forecast data from WeatherBench2 available through Google

Cloud. We express our gratitude to the researchers at ECMWF for providing the ERA5 dataset and HRES to the research community. We acknowledged the efforts of NOAA National Centers for Environmental Information in making the IBTrACS dataset available. The computations in this research were performed using the CFFF platform of Fudan University. We also thank support from the Computing for the Future at Fudan (CFFF). We are grateful to Shanghai Academy of Artificial Intelligence for Science (SAIS) for providing the internship opportunity, and this work was primarily completed during the internship period.

Author Contributions

Jun Liu designed the experiments, processed the data, performed evaluation analyses and calculations, and wrote the manuscript. Tao Zhou conducted the circulation field analysis. Jiarui Li wrote the evaluation metrics sections of the manuscript. Xiaohui Zhong, Peng Zhang and Jie Feng contributed to manuscript review. Xiaohui Zhong, Lei Chen, and Hao Li developed the FuXi-ENS model. Hao Li supervised the project and provided overall planning, management, and guidance.

Ethics declarations

The authors declare no competing interests.

References

- [1] Emanuel, K.: Increasing destructiveness of tropical cyclones over the past 30 years. *Nature* **436**(7051), 686–688 (2005)
- [2] Mendelsohn, R., Emanuel, K., Chonabayashi, S., Bakkensen, L.: The impact of climate change on global tropical cyclone damage. *Nature climate change* **2**(3), 205–209 (2012)
- [3] Peduzzi, P., Chatenoux, B., Dao, H., De Bono, A., Herold, C., Kossin, J., Mouton, F., Nordbeck, O.: Global trends in tropical cyclone risk. *Nature climate change* **2**(4), 289–294 (2012)
- [4] Zhang, Q., Wu, L., Liu, Q.: Tropical cyclone damages in china 1983–2006. *Bulletin of the American Meteorological Society* **90**(4), 489–496 (2009)
- [5] Doocy, S., Dick, A., Daniels, A., Kirsch, T.D.: The human impact of tropical cyclones: a historical review of events 1980–2009 and systematic literature review. *PLoS currents* **5**, (2013)

- [6] Qin, L., Zhu, L., Liao, X., Meng, C., Han, Q., Li, Z., Shen, S., Xu, W., Chen, J.: Recent northward shift of tropical cyclone economic risk in china. *npj Natural Hazards* **1**(1), 8 (2024)
- [7] Palmer, T.N.: Predicting uncertainty in forecasts of weather and climate. *Reports on progress in Physics* **63**(2), 71 (2000)
- [8] Lang, S., Alexe, M., Clare, M.C., Roberts, C., Adewoyin, R., Bouallègue, Z.B., Chantry, M., Dramsch, J., Dueben, P.D., Hahner, S., et al.: Aifs-crps: ensemble forecasting using a model trained with a loss function based on the continuous ranked probability score. *arXiv preprint arXiv:2412.15832* (2024)
- [9] Leutbecher, M., Palmer, T.N.: Ensemble forecasting. *Journal of computational physics* **227**(7), 3515–3539 (2008)
- [10] Feng, J., Toth, Z., Zhang, J., Peña, M.: Ensemble forecasting: a foray of dynamics into the realm of statistics. *Quarterly Journal of the Royal Meteorological Society* **150**(762), 2537–2560 (2024)
- [11] Wang, C., Zeng, Z., Ying, M.: Uncertainty in tropical cyclone intensity predictions due to uncertainty in initial conditions. *Advances in Atmospheric Sciences* **37**(3), 278–290 (2020)
- [12] Zhang, Q., Gu, X., Shi, P., Singh, V.P.: Impact of tropical cyclones on flood risk in southeastern china: Spatial patterns, causes and implications. *Global and Planetary Change* **150**, 81–93 (2017)
- [13] Buizza, R., Milleer, M., Palmer, T.N.: Stochastic representation of model uncertainties in the ecmwf ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society* **125**(560), 2887–2908 (1999)
- [14] Buizza, R.: Introduction to the special issue on “25 years of ensemble forecasting”. *Quarterly Journal of the Royal Meteorological Society* **145**, 1–11 (2019)
- [15] Buizza, R.: Localization of optimal perturbations using a projection operator. *Quarterly Journal of the Royal Meteorological Society* **120**(520), 1647–1681 (1994)
- [16] Buizza, R., Palmer, T.N.: The singular-vector structure of the atmospheric global circulation. *Journal of Atmospheric Sciences* **52**(9), 1434–1456 (1995)
- [17] Buizza, R., Leutbecher, M., Isaksen, L.: Potential use of an ensemble of analyses in the ecmwf ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences*,

- applied meteorology and physical oceanography **134**(637), 2051–2066 (2008)
- [18] Wastl, C., Wang, Y., Atencia, A., Wittmann, C.: Independent perturbations for physics parametrization tendencies in a convection-permitting ensemble (psppt). *Geoscientific Model Development* **12**(1), 261–273 (2019)
 - [19] Langland, R.H., Maue, R.N.: Recent northern hemisphere mid-latitude medium-range deterministic forecast skill. *Tellus A: Dynamic Meteorology and Oceanography* **64**(1), 17531 (2012)
 - [20] Molteni, F., Buizza, R., Palmer, T.N., Petroliagis, T.: The ecmwf ensemble prediction system: Methodology and validation. *Quarterly journal of the royal meteorological society* **122**(529), 73–119 (1996)
 - [21] Palmer, T.: The ecmwf ensemble prediction system: Looking back (more than) 25 years and projecting forward 25 years. *Quarterly Journal of the Royal Meteorological Society* **145**, 12–24 (2019)
 - [22] Aijaz, S., Kepert, J.D., Ye, H., Huang, Z., Hawksford, A.: Bias correction of tropical cyclone parameters in the ecmwf ensemble prediction system in australia. *Monthly Weather Review* **147**(11), 4261–4285 (2019)
 - [23] Elsberry, R.L., Tsai, H.-C., Chin, W.-C., Marchok, T.P.: Predicting rapid intensification events following tropical cyclone formation in the western north pacific based on ecmwf ensemble warm core evolutions. *Atmosphere* **12**(7), 847 (2021)
 - [24] Hamill, T.M., Bates, G.T., Whitaker, J.S., Murray, D.R., Fiorino, M., Galarneau Jr, T.J., Zhu, Y., Lapenta, W.: Noaa’s second-generation global medium-range ensemble reforecast dataset. *Bulletin of the American Meteorological Society* **94**(10), 1553–1565 (2013)
 - [25] Houtekamer, P.L., Mitchell, H.L.: A sequential ensemble kalman filter for atmospheric data assimilation. *Monthly weather review* **129**(1), 123–137 (2001)
 - [26] Bishop, C.H., Etherton, B.J., Majumdar, S.J.: Adaptive sampling with the ensemble transform kalman filter. part i: Theoretical aspects. *Monthly weather review* **129**(3), 420–436 (2001)
 - [27] Zhang, F., Weng, Y., Sippel, J.A., Meng, Z., Bishop, C.H.: Cloud-resolving hurricane initialization and prediction through assimilation of doppler radar observations with an ensemble kalman filter. *Monthly Weather Review* **137**(7), 2105–2125 (2009)

- [28] Zhou, L., Lin, S.-J., Chen, J.-H., Harris, L.M., Chen, X., Rees, S.L.: Toward convective-scale prediction within the next generation global prediction system. *Bulletin of the American Meteorological Society* **100**(7), 1225–1243 (2019)
- [29] Hazelton, A.T., Harris, L., Lin, S.-J.: Evaluation of tropical cyclone structure forecasts in a high-resolution version of the multiscale gfdl fvgs model. *Weather and Forecasting* **33**(2), 419–442 (2018)
- [30] Zhang, F., Zhang, M., Hansen, J.A.: Coupling ensemble kalman filter with four-dimensional variational data assimilation. *Advances in Atmospheric Sciences* **26**(1), 1–8 (2009)
- [31] Zhang, M., Zhang, F.: E4dvar: Coupling an ensemble kalman filter with four-dimensional variational data assimilation in a limited-area weather prediction model. *Monthly Weather Review* **140**(2), 587–600 (2012)
- [32] Buizza, R., Houtekamer, P., Pellerin, G., Toth, Z., Zhu, Y., Wei, M.: A comparison of the ecmwf, msc, and ncep global ensemble prediction systems. *Monthly Weather Review* **133**(5), 1076–1097 (2005)
- [33] Bauer, P., Thorpe, A., Brunet, G.: The quiet revolution of numerical weather prediction. *Nature* **525**(7567), 47–55 (2015)
- [34] Pu, J., Mu, M., Feng, J., Zhong, X., Li, H.: A fast physics-based perturbation generator of machine learning weather model for efficient ensemble forecasts of tropical cyclone track. *npj Climate and Atmospheric Science* **8**(1), 128 (2025)
- [35] Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z., Azizzadenesheli, K., et al.: Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214* (2022)
- [36] Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., Tian, Q.: Pangu-weather: A 3d high-resolution model for fast and accurate global weather forecast. *arXiv preprint arXiv:2211.02556* (2022)
- [37] Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., et al.: Learning skillful medium-range global weather forecasting. *Science* **382**(6677), 1416–1421 (2023)
- [38] Chen, L., Zhong, X., Zhang, F., Cheng, Y., Xu, Y., Qi, Y., Li, H.: Fuxi: a cascade machine learning forecasting system for 15-day global weather forecast. *npj climate and atmospheric science* **6**(1), 190 (2023)

- [39] Lang, S., Leutbecher, M., Maciel, P.: A multi-scale loss formulation for learning a probabilistic model with proper score optimisation. arXiv preprint arXiv:2506.10868 (2025)
- [40] Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T.R., El-Kadi, A., Masters, D., Ewalds, T., Stott, J., Mohamed, S., Battaglia, P., *et al.*: Probabilistic weather forecasting with machine learning. *Nature* **637**(8044), 84–90 (2025)
- [41] Zhong, X., Chen, L., Li, H., Liu, J., Fan, X., Feng, J., Dai, K., Luo, J.-J., Wu, J., Lu, B.: Fuxi-ens: A machine learning model for medium-range ensemble weather forecasting. arXiv preprint arXiv:2405.05925 (2024)
- [42] Bonavita, M., Hólm, E., Isaksen, L., Fisher, M.: The evolution of the ecmwf hybrid data assimilation system. *Quarterly Journal of the Royal Meteorological Society* **142**(694), 287–303 (2016)
- [43] Houtekamer, P.L., Zhang, F.: Review of the ensemble kalman filter for atmospheric data assimilation. *Monthly Weather Review* **144**(12), 4489–4532 (2016)
- [44] Palmer, T., Gelaro, R., Barkmeijer, J., Buizza, R.: Singular vectors, metrics, and adaptive observations. *Journal of the Atmospheric Sciences* **55**(4), 633–653 (1998)
- [45] Ehrendorfer, M., Errico, R.M., Raeder, K.D.: Singular-vector perturbation growth in a primitive equation model with moist physics. *Journal of the Atmospheric Sciences* **56**(11), 1627–1648 (1999)
- [46] Gray, W.M.: Global view of the origin of tropical disturbances and storms. *Monthly Weather Review* **96**(10), 669–700 (1968)
- [47] Emanuel, K.A.: An air-sea interaction theory for tropical cyclones. part i: Steady-state maintenance. *Journal of Atmospheric Sciences* **43**(6), 585–605 (1986)
- [48] Done, J.M., Lackmann, G.M., Prein, A.F.: The response of tropical cyclone intensity to changes in environmental temperature. *Weather and Climate Dynamics* **3**(2), 693–711 (2022)
- [49] Richardson, D.S.: Skill and relative economic value of the ecmwf ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society* **126**(563), 649–667 (2000)
- [50] Dulac, W., Cattiaux, J., Chauvin, F., Bourdin, S., Fromang, S.: Assessing the representation of tropical cyclones in era5 with the cnrm tracker. *Climate Dynamics* **62**(1), 223–238 (2024)

- [51] Liu, C.-C., Hsu, K., Peng, M.S., Chen, D.-S., Chang, P.-L., Hsiao, L.-F., Fong, C.-T., Hong, J.-S., Cheng, C.-P., Lu, K.-C., *et al.*: Evaluation of five global ai models for predicting weather in eastern asia and western pacific. *npj Climate and Atmospheric Science* **7**(1), 221 (2024)
- [52] Subich, C., Husain, S.Z., Separovic, L., Yang, J.: Fixing the double penalty in data-driven weather forecasting through a modified spherical harmonic loss function. *arXiv preprint arXiv:2501.19374* (2025)
- [53] Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., *et al.*: The era5 global reanalysis. *Quarterly journal of the royal meteorological society* **146**(730), 1999–2049 (2020)
- [54] 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. *Bulletin of the American Meteorological Society* **91**(3), 363–376 (2010)
- [55] Kenneth, R., Howard, J., James, P., Michael, C., Carl, J.: International best track archive for climate stewardship (ibtracs) project, version 4. (No Title) (2019)
- [56] Kingma, D.P., Welling, M.: Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013)
- [57] Kingma, D.P., Welling, M., *et al.*: An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning* **12**(4), 307–392 (2019)
- [58] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 9992–10002 (2021). <https://doi.org/10.1109/iccv48922.2021.00986>
- [59] Zhong, X., Chen, L., Liu, J., Lin, C., Qi, Y., Li, H.: Fuxi-extreme: Improving extreme rainfall and wind forecasts with diffusion model. *Science China Earth Sciences* **67**(12), 3696–3708 (2024)
- [60] Zhang, X., Chen, G., Yu, H., Zeng, Z.: Verification of ensemble track forecasts of tropical cyclones during 2014. *Tropical Cyclone Research and Review* **4**(2), 79–87 (2015)
- [61] Mason, S.J., Graham, N.E.: Areas beneath the relative operating characteristics (roc) and relative operating levels (rol) curves: Statistical significance and interpretation. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology*

- and physical oceanography **128**(584), 2145–2166 (2002)
- [62] Whitaker, J.S., Loughe, A.F.: The relationship between ensemble spread and ensemble mean skill. *Monthly Weather Review* **126**(12), 3292–3302 (1998)
 - [63] Yamaguchi, M., Sakai, R., Kyoda, M., Komori, T., Kadowaki, T.: Typhoon ensemble prediction system developed at the japan meteorological agency. *Monthly Weather Review* **137**(8), 2592–2604 (2009)
 - [64] Goerss, J.S.: Tropical cyclone track forecasts using an ensemble of dynamical models. *Monthly Weather Review* **128**(4), 1187–1193 (2000)
 - [65] Aberson, S.D.: The ensemble of tropical cyclone track forecasting models in the north atlantic basin (1976–2000). *Bulletin of the American Meteorological Society* **82**(9), 1895–1904 (2001)
 - [66] Cangialosi, J.P., Franklin, J.: National hurricane center forecast verification report. National Hurricane Center (NHC) **79** (2012)
 - [67] Goerss, J.S., Sampson, C.R., Gross, J.M.: A history of western north pacific tropical cyclone track forecast skill. *Weather and Forecasting* **19**(3), 633–638 (2004)
 - [68] Kirkland, J.L., Zick, S.E.: Regional differences in the spatial patterns of north atlantic tropical cyclone rainbands through landfall. *southeastern geographer* **59**(3), 294–320 (2019)
 - [69] Nachar, N., *et al.*: The mann-whitney u: A test for assessing whether two independent samples come from the same distribution. *Tutorials in quantitative Methods for Psychology* **4**(1), 13–20 (2008)
 - [70] Titley, H.A., Bowyer, R.L., Cloke, H.L.: A global evaluation of multi-model ensemble tropical cyclone track probability forecasts. *Quarterly Journal of the Royal Meteorological Society* **146**(726), 531–545 (2020). <https://doi.org/10.1002/qj.3712>
 - [71] Gentile, E.S., Gray, S.L., Lewis, H.W.: The sensitivity of probabilistic convective-scale forecasts of an extratropical cyclone to atmosphere–ocean–wave coupling. *Quarterly Journal of the Royal Meteorological Society* **148**(743), 685–710 (2022). <https://doi.org/10.1002/qj.4225>
 - [72] Shao, D., Zhang, Y., Xu, J., *et al.*: Impact of model perturbation schemes on regional ensemble forecasting of typhoons based on typhoon kompasu (2118). *Chinese Journal of Atmospheric Sciences* **48**(4), 1511–1530 (2024). <https://doi.org/10.3878/j.issn.1006-9895.2301.22122>

Supplementary information

Contents of this file

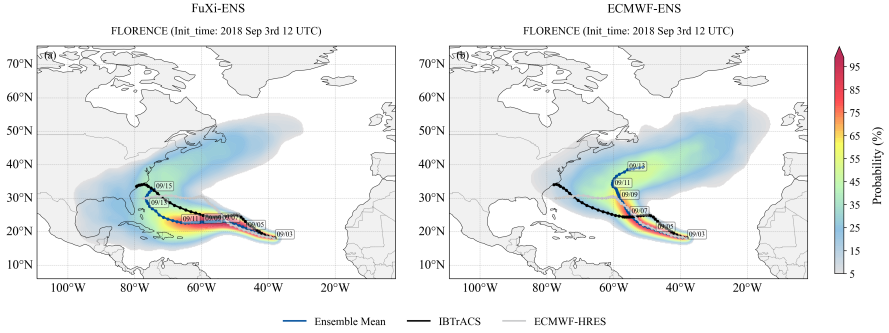
Supplementary Table 1
Supplementary Figures 1 to 11

Supplementary Table

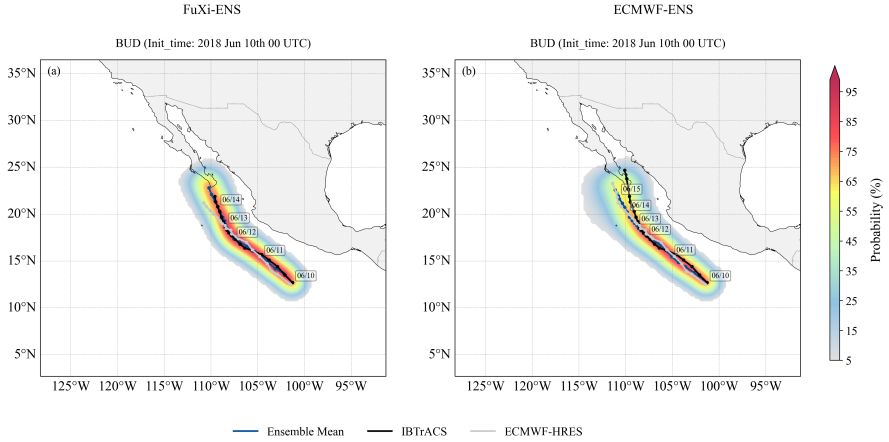
Table 1: Parameters and thresholds related to the TC tracking.

Parameter	Level	Radius	Threshold	Comments
WS10M	Surface	278 km	$> 8 \text{ m s}^{-1}$	Only required over land.
Vorticity	850 hPa	278 km	absolute value $\geq 5 \times 10^{-5} \text{ s}^{-1}$	
Geopotential thickness	850 hPa and 200 hPa	278 km		A maximum in thickness is necessary only after the TC has transitioned into an extra-tropical cyclone.

Supplementary Figures

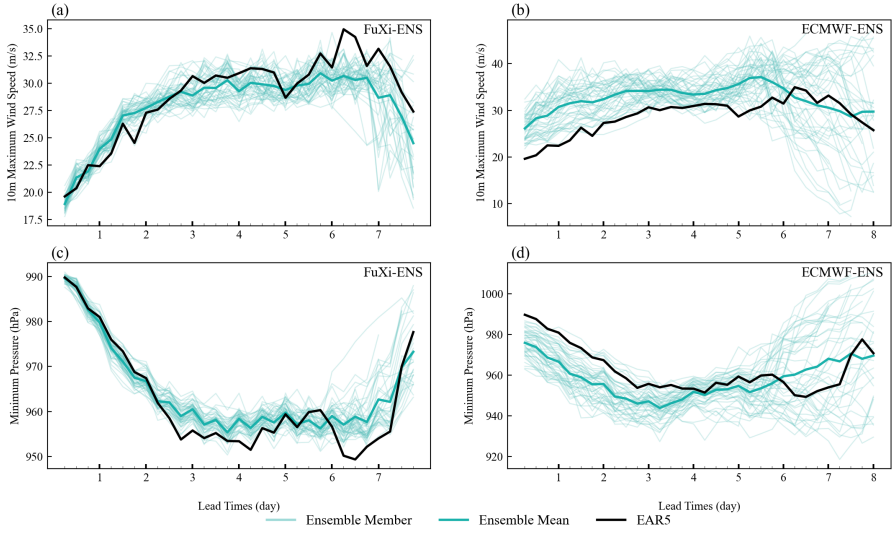


Supplementary Fig. 1: Probabilistic forecast comparison for hurricane FLORENCE track prediction. Hurricane FLORENCE track probabilistic forecast distribution based on initialization at 12 UTC on September 3, 2018: (a) FuXi-ENS and (b) ECMWF-ENS. Blue lines represent ensemble mean tracks, black lines represent IBTrACS observed tracks, and gray lines represent ECMWF-HRES (deterministic forecast). Shading indicates TC strike probability (%). Date labels mark the temporal evolution of the track.

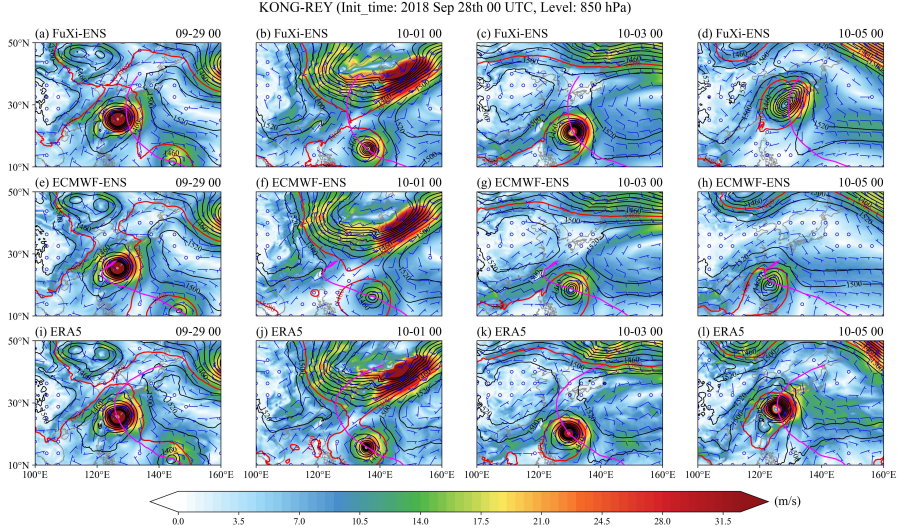


Supplementary Fig. 2: Probabilistic forecast comparison for hurricane BUD track prediction. Hurricane BUD track probabilistic forecast distribution based on initialization at 00 UTC on June 10, 2018: (a) FuXi-ENS and (b) ECMWF-ENS. Blue lines represent ensemble mean tracks, black lines represent IBTrACS observed tracks, and gray lines represent ECMWF-HRES (deterministic forecast). Shading indicates TC strike probability (%). Date labels mark the temporal evolution of the track.

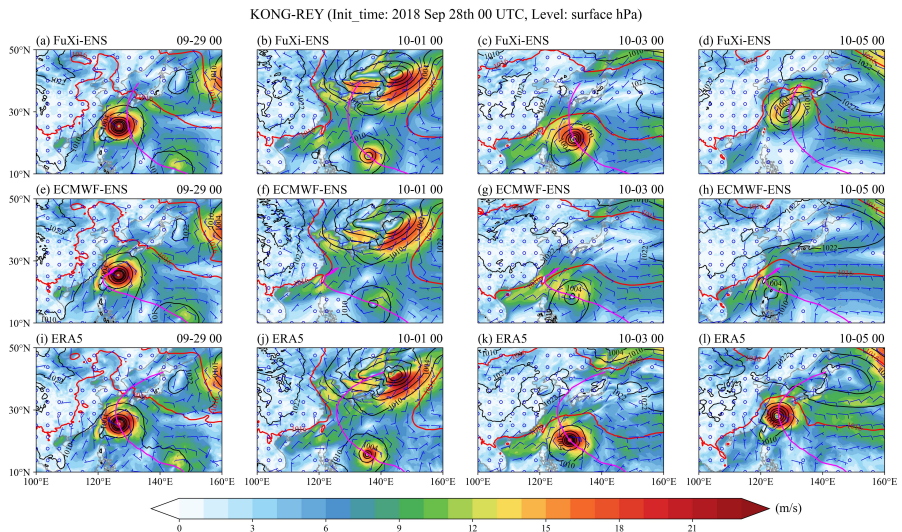
TRAMI
Init_time: 2018 Sep 23rd 00 UTC



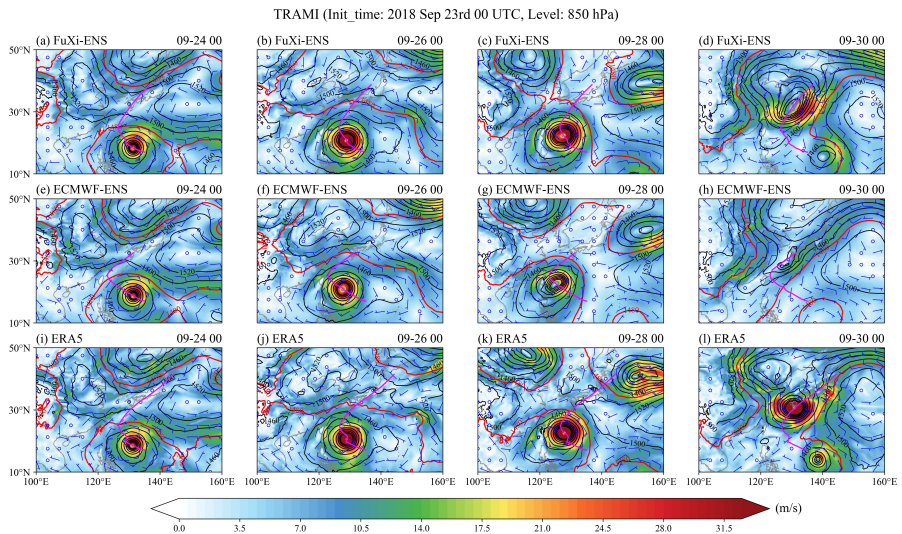
Supplementary Fig. 3: Intensity forecast comparison for typhoon TRAMI. 10m maximum wind speed variation with forecast lead time for (a) FuXi-ENS and (b) ECMWF-ENS initialized at 00 UTC on September 23, 2018. Minimum pressure variation with forecast lead time for (c) FuXi-ENS and (d) ECMWF-ENS. Light green thin lines represent ensemble member forecasts, dark green thick lines represent ensemble mean forecasts, and black thick lines represent ERA5 reanalysis.



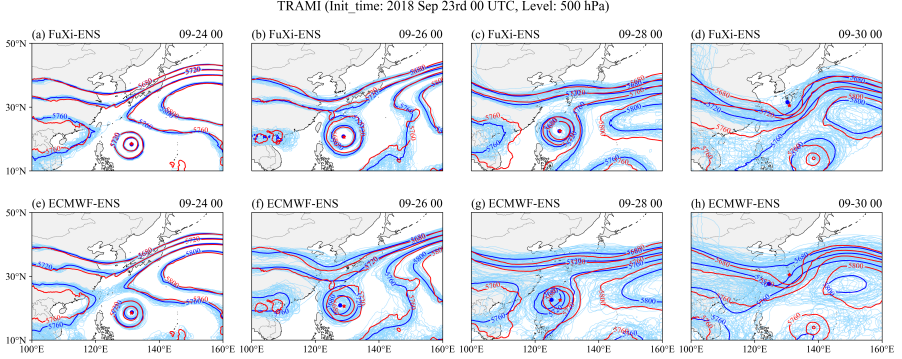
Supplementary Fig. 4: Evolution of 850 hPa atmospheric circulation patterns for typhoon KONG-REY with forecast lead time. Evolution of KONG-REY's 850 hPa wind speed (shading, m/s), geopotential height (contours, gpm, 20 gpm intervals), and horizontal wind (vectors, m/s) initialized at 00 UTC on September 28, 2018, for September 29 (first column), October 1 (second column), October 3 (third column), and October 5 (fourth column) from FuXi-ENS (first row), ECMWF-ENS (second row), and ERA5 (third row). Red contours indicate the 1480 gpm contour line, and magenta lines denote the best track (IBTrACS).



Supplementary Fig. 5: Evolution of surface atmospheric circulation patterns for typhoon KONG-REY with forecast lead time. Evolution of KONG-REY's 10m wind speed (shading, m/s), mean sea level pressure (contours, hPa, 6 hPa intervals), and 10m wind (vectors, m/s) initialized at 00 UTC on September 28, 2018, for September 29 (first column), October 1 (second column), October 3 (third column), and October 5 (fourth column) from FuXi-ENS (first row), ECMWF-ENS (second row), and ERA5 (third row). Red contours indicate the 1016 isoline, and magenta lines denote the best track (IBTrACS).

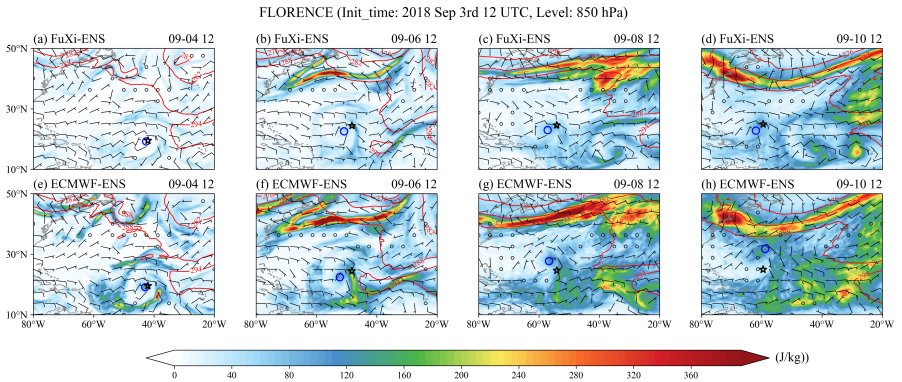


Supplementary Fig. 6: Evolution of 850 hPa atmospheric circulation patterns for typhoon TRAMI with forecast lead time. Evolution of TRAMI's 850 hPa wind speed (shading, m/s), geopotential height (contours, gpm, 20 gpm intervals), and horizontal wind (vectors, m/s) initialized at 00 UTC on September 23, 2018, for September 24 (first column), September 26 (second column), September 28 (third column), and September 30 (fourth column) from FuXi-ENS (first row), ECMWF-ENS (second row), and ERA5 (third row). Red contours indicate the 1480 gpm contour line, and magenta lines denote the best track (IBTrACS).

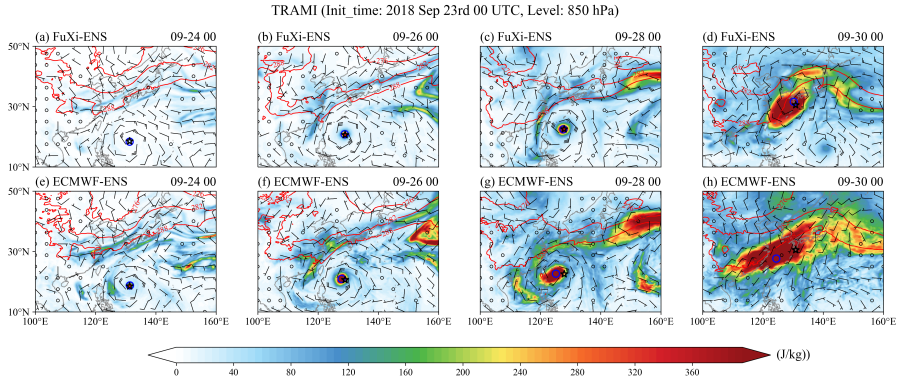


Supplementary Fig. 7: Evolution of 500 hPa geopotential height ensemble forecasts for typhoon TRAMI with forecast lead time.

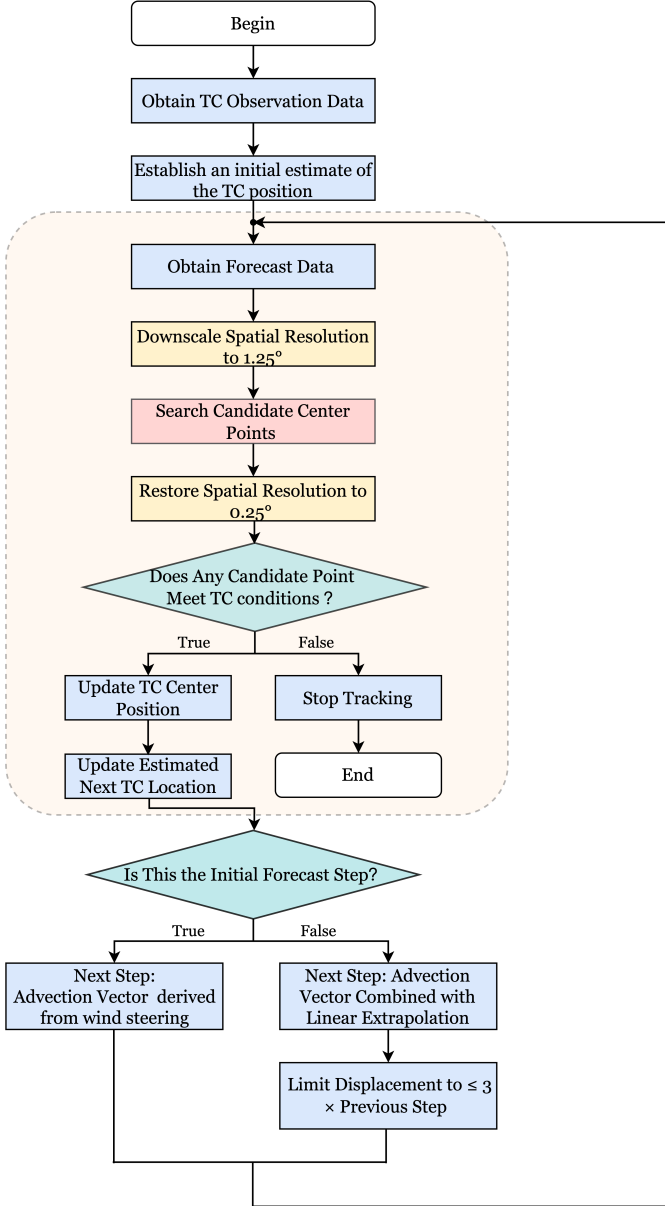
Evolution of TRAMI's 500 hPa geopotential height (contours, gpm) initialized at 00 UTC on September 23, 2018, for September 24 (first column), September 26 (second column), September 28 (third column), and September 30 (fourth column) from FuXi-ENS (first row), ECMWF-ENS (second row). Light blue lines represent individual ensemble members, dark blue lines represent the ensemble mean, and red lines represent ERA5 reanalysis. Four contour lines are shown: 5680 gpm, 5720 gpm, 5760 gpm and 5800 gpm. Blue dots denote the ensemble mean track and red stars denote the best track (IBTrACS).



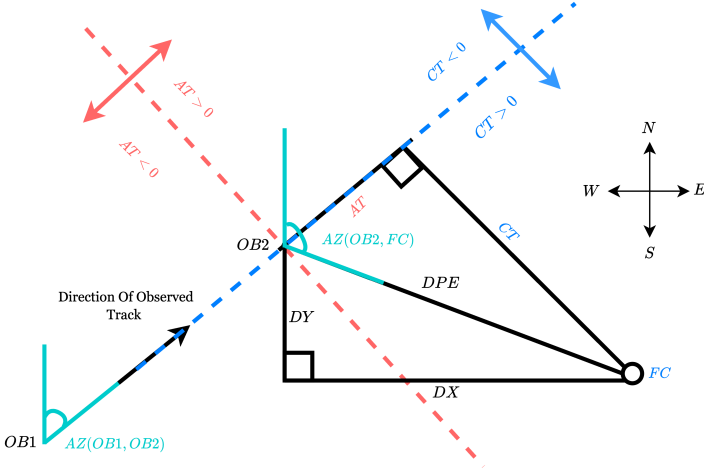
Supplementary Fig. 8: Evolution of 850 hPa MTE for hurricane FLORENCE with forecast lead time. Evolution of FLORENCE's 850 hPa MTE (shading, J/kg), temperature (contours, K), and horizontal wind (vectors, m/s) initialized at 12 UTC on September 3, 2018, for September 4 (first column), September 6 (second column), September 8 (third column), and September 10 (fourth column) from FuXi-ENS (first row), ECMWF-ENS (second row). Blue circles denote the ensemble mean track and black stars denote the best track (IBTrACS).



Supplementary Fig. 9: Evolution of 850 hPa MTE for typhoon TRAMI with forecast lead time. Evolution of TRAMI's 850 hPa MTE (shading, J/kg), temperature (contours, K), and horizontal wind (vectors, m/s) initialized at 00 UTC on September 28, 2018, for September 29 (first column), October 1 (second column), October 3 (third column), and October 5 (fourth column) from FuXi-ENS (first row), ECMWF-ENS (second row). Blue circles denote the ensemble mean track and black stars denote the best track (IBTrACS).



Supplementary Fig. 10: Schematic diagram of the TC ensemble forecast tracking algorithm. The algorithm obtains TC observation data to establish an initial estimate of the TC position, then obtains forecast data for downscaling spatial resolution and searching candidate center points. TC center position is updated when any candidate point meets TC conditions, otherwise stop tracking. Subsequent steps use advection vector derived from wind steering or advection vector combined with linear extrapolation, depending on whether this is the initial forecast step.



Supplementary Fig. 11: Geometric schematic diagram of TC track forecast error decomposition. This schematic diagram illustrates the calculation method for along-track error (AT) and cross-track error (CT) in TC track forecasting. OB_1 and OB_2 represent the TC center positions at two consecutive observation times, and FC represents the forecast position at the corresponding time. $AZ(OB_1, OB_2)$ is the azimuth angle of the observed track, defining the actual movement direction of the TC; $AZ(OB_2, FC)$ is the azimuth angle from the latest observed position to the forecast position. DPE (Direct Position Error) represents the straight-line distance between the observed and forecast positions, which can be decomposed into horizontal component DX and vertical component DY . AT is defined as the projection error of the forecast position in the direction of the observed track: $AT > 0$ indicates the forecast position is ahead of the observed track, $AT < 0$ indicates the forecast position lags behind the observed track. CT is defined as the deviation of the forecast position perpendicular to the observed track direction: $CT > 0$ indicates the forecast position is on the right side of the observed track, $CT < 0$ indicates the forecast position is on the left side of the observed track. The red and blue dashed lines respectively mark the positive and negative regional divisions of AT and CT. The compass in the upper right corner indicates the geographic coordinate system (N-North, S-South, E-East, W-West). This decomposition method helps quantitatively analyze the systematic bias characteristics in TC track forecasting.