Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Can more accurate atmospheric forecasts translate into improved tropical cyclone intensity prediction? A comparison of AIFS and IFS

Hongyu Qu
State Key Laboratory of Severe Weather Meteorological Science and Technology, Chinese Academy of Meteorological Sciences, China Meteorological Administration

Hongxiong Xu
xuhx@cma.gov.cn

State Key Laboratory of Severe Weather Meteorological Science and Technology, Chinese Academy of Meteorological Sciences, China Meteorological Administration

Lin Dong
National Meteorological Centre

Chunyi Xiang
National Meteorological Centre

Dajun Zhao
State Key Laboratory of Severe Weather Meteorological Science and Technology, Chinese Academy of Meteorological Sciences, China Meteorological Administration

Article

Keywords:

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11521313/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Can more accurate atmospheric forecasts translate into improved tropical cyclone intensity prediction? A comparison of AIFS and IFS

Hongyu Qu[1,2], Hongxiong Xu*[1], Lin Dong[2], Chunyi Xiang[2], and Dajun Zhao[1]

[1]State Key Laboratory of Severe Weather Meteorological Science and Technology, Chinese Academy of Meteorological Sciences, China Meteorological Administration, Beijing, China
[2]National Meteorological Centre, Beijing, China

Abstract

Advances in numerical weather prediction have substantially improved large-scale atmospheric modeling and tropical cyclone (TC) track forecasts, but gains in TC intensity prediction have been more limited. A similar gap persists for AI weather models, which can predict atmospheric evolution with high skill while remaining less skillful than numerical models in direct prediction of TC intensity. Whether their advantage in atmospheric forecasting can be converted into an improved TC intensity forecast remains unclear. Here, we compare the Artificial Intelligence Forecasting System (AIFS) and the Integrated Forecasting System (IFS), both as direct TC forecasting systems and as atmospheric inputs to the Tropical cyclone Intensity Forecasting Network (TIFNet), a cyclone-specific AI model for TC intensity prediction. AIFS has larger direct intensity errors than IFS but lower atmospheric-field errors from 24 to 120 h. When used as TIFNet inputs, however, AIFS forecasts produce lower intensity errors from 72 to 120 h, with the largest improvements for super typhoons and storms undergoing rapid changes in intensity. Controlled sensitivity experiments show that lower atmospheric-field errors offset part of the loss associated with reduced short-wavelength variability, while input-specific fine-tuning recovers additional skill. The findings of this study show that more accurate environmental prediction can translate into improved forecasts of TC intensity despite smoother storm-scale representation, thereby providing a practical route for combining global AI weather forecasts with specialized TC intensity models.

1 Introduction

Tropical cyclones (TCs) are among the most destructive weather systems, making accurate forecasts of their intensity essential for anticipating their impact (Peduzzi et al., 2012; Zhang et al., 2009). However, TC intensity remains difficult to predict because it depends on the coupled evolution of the vortex itself and its surrounding environment. Storm structure, convection, and boundary layer dynamics shape the evolution of the vortex, while thermodynamic conditions, vertical wind shear, and air–sea exchange regulate storm intensity (DeMaria, 1996; Emanuel, 1999; Jiang, 2012; Emanuel and Zhang, 2016; Montgomery and Smith, 2017). Errors in simulating either the vortex or its environment can therefore propagate into subsequent modeling of the evolution of storm intensity, particularly during periods of rapid intensification (RI) or rapid weakening (RW) (Kaplan et al., 2010; DeMaria et al., 2014; Cangialosi et al., 2020). Consequently, conventional dynamical TC prediction depends on accurate representation of both the evolving storm and the regional environment that governs its intensity change.

*Correspondence to Hongxiong Xu, xuhx@cma.gov.cn

Over recent decades, advances in numerical weather prediction have substantially improved large-scale atmospheric forecasts and TC track prediction, whereas progress in TC intensity prediction has been more modest. Higher model resolution, improved data assimilation, and advances in physical parameterizations have strengthened the representation of TCs and their interactions with the surrounding environment (Bauer et al., 2015; Gopalakrishnan et al., 2012; Tallapragada et al., 2014). Nevertheless, TC intensity forecasts remain sensitive to uncertainties in the initial vortex, model physics, and the limited predictability of the inner-core evolution (Emanuel and Zhang, 2016). Consequently, gains in atmospheric forecast skill have not translated into comparable gains in TC intensity prediction. The matter of whether increasingly accurate atmospheric forecasts can be translated into improved TC intensity forecasts remains to be resolved.

The rapid development of AI weather prediction makes this problem even more important. Global AI weather models now achieve high skill in atmospheric forecasting at substantially lower computational cost than conventional dynamical systems (Bi et al., 2023; Lam et al., 2023; Price et al., 2025). However, improvements in global forecast skill do not necessarily translate into improved direct forecasts of TC intensity. Evaluations of global AI weather models have found intensity errors that can remain larger than those of numerical weather prediction systems (DeMaria et al., 2025; Yamaguchi et al., 2025). Their forecasts also tend to contain smoother cyclone structures and less small-scale variability, which might limit the representation of storm-scale processes relevant to TC intensity evolution (Bi et al., 2023; Bonavita, 2024). Moreover, accurate prediction of the evolving atmospheric environment might retain information about subsequent TC intensity change, even with smoother representation of the storm itself. Whether this information can be converted into improved TC intensity forecasts is a separate question to how accurately global models predict TC intensity directly.

Answering this question requires a model that can reliably translate environmental forecast information into TC intensity forecasts. Statistical and machine learning intensity models provide such a link by relating atmospheric predictors directly to subsequent TC intensity evolution, rather than relying on the intensity simulated by the atmospheric model itself (DeMaria et al., 2005; Xu et al., 2021). Recent studies have reported improved TC intensity forecasts using models that combine storm information with atmospheric fields (Lee et al., 2024; Huang et al., 2025; Wang et al., 2025). The Tropical cyclone Intensity Forecasting Network (TIFNet) combines forecast atmospheric fields with recent TC intensity evolution and directly predicts multi-lead intensity changes using a noniterative spatiotemporal Transformer. It has demonstrated strong skill relative to operational guidance, particularly during periods of rapid TC intensification (Qu et al., 2026). More recent work has independently shown that substantial TC intensity information can also be extracted from relatively coarse atmospheric fields, further supporting the potential of cyclone-specific AI prediction based on global atmospheric forecasts (Alet et al., 2026). Applying the same intensity model to forecasts from different atmospheric systems provides a controlled test of whether differences in atmospheric forecast accuracy translate into differences in TC intensity prediction.

Here, we compare the European Centre for Medium-Range Weather Forecasts Integrated Forecasting System (IFS) and the Artificial Intelligence Forecasting System (AIFS), both as direct TC forecasting systems and as atmospheric inputs to TIFNet. By applying the same intensity model to both sets of forecast fields, we assess whether differences in environmental forecast accuracy translate into differences in TC intensity prediction. We then use controlled experiments based on forecast-error growth, spectral perturbation, and input-specific fine-tuning to separate the effects of atmospheric forecast accuracy, short-wavelength information, and model adaptation. This analysis is intended to clarify how advances in global atmospheric prediction can be translated into improved forecasts of TC intensity, and to establish a practical approach for combining global AI weather forecasts with specialized intensity prediction models.

2 Results

2.1 Tropical cyclone intensity forecasts

We first tested whether direct TC intensity skill reflects the usefulness of a model's atmospheric forecast fields for subsequent intensity prediction. AIFS and IFS were evaluated over the 2023–2024 TC seasons (Fig. 1a), first as direct forecasting systems and then as atmospheric inputs to TIFNet. TIFNet was fine-tuned separately for the two forecast sources, yielding TIFNet-AIFS and TIFNet-IFS. All forecasts were verified against China Meteorological Administration (CMA) best-track data (Ying et al., 2014; Lu et al., 2021).

The direct forecasts gave contrasting results for intensity and track. AIFS had larger intensity errors than IFS at every evaluated lead time, with differences in mean absolute error (MAE) of 5.21–7.33 $m\,s^{-1}$ (Fig. 1b). In contrast, AIFS had smaller track errors throughout the same period, with reductions of 3.88–79.62 km (Fig. 1c). Differences in both intensity and track error were statistically significant at every evaluated lead time.

The intensity ranking changed when the atmospheric forecasts were used as inputs to TIFNet. TIFNet-AIFS and TIFNet-IFS had nearly identical MAEs at 12 h, with only small differences from 24 to 60 h (Fig. 1d). From 72 to 120 h, TIFNet-AIFS consistently had lower MAEs, with reductions of 0.06–1.55 $m\,s^{-1}$. The largest reduction occurred at 108 h, when the TIFNet-AIFS MAE was 1.55 $m\,s^{-1}$ (18.71%) lower than that of TIFNet-IFS.

The difference was larger during periods of rapid change in intensity. For RI and RW samples, TIFNet-AIFS had lower MAEs at every evaluated lead time, with reductions of 0.05–2.22 $m\,s^{-1}$ (Fig. 1e). The largest reduction occurred at 96 h and reached 2.22 $m\,s^{-1}$ (27.60%). Thus, the direct intensity ranking of AIFS and IFS was not preserved when their atmospheric fields were used as inputs to the same TIFNet architecture with input-specific fine-tuning. The two TIFNet configurations performed similarly at shorter lead times, whereas TIFNet-AIFS had lower errors at longer lead times and during rapid changes in TC intensity.

2.2 Forecast skill across intensity categories

The difference between TIFNet-AIFS and TIFNet-IFS depended strongly on TC intensity (Fig. 2). For tropical storms (TS), TIFNet-AIFS had numerically lower MAEs from 48 to 120 h (Fig. 2a), whereas differences for severe tropical storms (STS) and severe typhoons (STY) were small and changed sign across lead times (Fig. 2b,d). For typhoons (TY), TIFNet-AIFS had higher MAEs at 48 and 72 h by 1.23 and 1.72 $m\,s^{-1}$, respectively (Fig. 2c). The largest and most persistent differences occurred for super typhoons (SuperTY): TIFNet-AIFS reduced the MAEs by 2.06, 3.27, and 2.75 $m\,s^{-1}$ at 72, 96, and 120 h, corresponding to reductions of 23.66%, 31.34%, and 22.25%, respectively (Fig. 2e). The advantage of TIFNet-AIFS was most pronounced for SuperTY.

Two SuperTY cases with contrasting intensity histories, Mawar (2023) and Man-yi (2024), were examined to illustrate the forecast differences during typhoon evolution (Figs. 3 and 4). Mawar underwent RI, brief weakening, and subsequent reintensification to SuperTY strength, whereas Man-yi underwent a shorter and more concentrated intensification–weakening cycle. TIFNet-AIFS and TIFNet-IFS had similar overall intensity errors at shorter lead times, but the difference increased at 96 and 120 h. The reductions were larger during rapid changes in intensity, reaching 4.58 $m\,s^{-1}$ at 72 h (Fig. 3f,g). For Man-yi, TIFNet-AIFS likewise produced lower MAEs for rapid-intensity-change samples at all five lead times, with reductions of 0.81–6.52 $m\,s^{-1}$ (Fig. 4g). Both cases were consistent in terms of the overall statistics, with larger differences between TIFNet-AIFS and TIFNet-IFS during periods of rapid change in intensity.

2.3 Accuracy of atmospheric forecast fields

To examine whether the longer-range differences between TIFNet-AIFS and TIFNet-IFS were associated with the accuracy of their atmospheric inputs, we compared AIFS and IFS forecast fields with verifying IFS analyses (Fig. 5). Root mean square errors (RMSEs) were evaluated for 16 surface and upper-air variable–level combinations. At 6 h, IFS had lower RMSEs than AIFS for all 16 fields. The ordering began to reverse during 12–18 h and AIFS had lower RMSEs from 24 to 120 h for every field examined. For most variables, the separation between the two systems increased with forecast lead time.

This ordering was the opposite to that of the direct intensity forecasts. IFS had lower direct intensity errors throughout the evaluation period (Fig. 1b), whereas AIFS had lower atmospheric-field errors from 24 to 120 h. The atmospheric-field advantage of AIFS emerged earlier and persisted throughout the 72–120-h forecast period, during which TIFNet-AIFS produced lower intensity errors than TIFNet-IFS. Direct intensity skill and atmospheric-field accuracy therefore did not produce the same ranking of the two forecast systems. Whether the lower atmospheric-field errors contributed to the TIFNet-AIFS advantage was examined through the controlled experiments described below.

2.4 Factors controlling TIFNet intensity forecasts

2.4.1 Sensitivity to input forecast-error growth

To test the sensitivity of TIFNet to atmospheric forecast error, we held the TIFNet parameters fixed and varied the forecast lead of the IFS input fields (Fig. 6). TIFNet-IFS-24h, TIFNet-IFS-48h, TIFNet-IFS-72h, and TIFNet-IFS-96h used IFS fields from the 24-, 48-, 72- and 96-h forecasts, respectively, with the same ERA5-pretrained TIFNet checkpoint applied without fine-tuning.

Across all samples, intensity MAEs increased with the forecast lead time of the input fields. Compared with TIFNet-IFS-24h, TIFNet-IFS-96h increased the MAEs by 2.15–3.31 $m\,s^{-1}$ across the evaluated TIFNet forecast lead times (Fig. 6a). The rapid-intensity-change samples showed the same overall tendency. Using 96-h rather than 24-h IFS fields increased the MAEs by 2.41–4.88 $m\,s^{-1}$, although the MAEs did not increase monotonically across the intermediate input lead times at every TIFNet forecast lead time (Fig. 6b). Thus, using IFS fields from longer forecasts was associated with less accurate TIFNet intensity forecasts, consistent with sensitivity to input error growth.

These results show that TIFNet intensity forecasts are strongly sensitive to the quality of the upstream atmospheric fields. As forecast errors grow with input lead time, the corresponding intensity forecasts deteriorate substantially. This sensitivity helps explain why the atmospheric-field accuracy of the upstream model matters for TIFNet, but it does not fully account for the differences between TIFNet-AIFS and TIFNet-IFS. We further examined whether the spatial structure of the input fields also affected intensity prediction.

2.4.2 Sensitivity to spectral information

AIFS and IFS also differed in the spatial spectra of their atmospheric fields. To isolate this difference, we constructed IFSspec by attenuating the Fourier amplitudes of the dynamic IFS fields towards the corresponding AIFS spectra while retaining the IFS spatial mean, Fourier phase, grid, and static variables. The 850-hPa kinetic energy spectra confirmed that the largest differences between IFS and AIFS occurred at shorter wavelengths and that IFSspec substantially reduced this excess short-wavelength power (Appendix Fig. A1).

For all forecast samples, spectral attenuation had little effect on intensity error: the MAE difference between TIFNet-IFSspec-ZS and TIFNet-IFS-ZS (ZS denotes zero-shot application of the ERA5-pretrained TIFNet without fine-tuning; the experimental configurations are summarized

in Table 1.) remained below 0.29 $m\,s^{-1}$ at every lead time (Fig. 7a). The response became much larger when forecasts were stratified by TC intensity. Spectral attenuation reduced the MAEs for TS and STS at all evaluated lead times but increased errors for stronger storms. For STY, the increase occurred from 48 to 120 h; for SuperTY, TIFNet-IFSspec-ZS had larger MAEs at every lead time, with increases of 0.62–2.43 $m\,s^{-1}$ (Fig. 7b–f).

Opposite responses occurred during RI and RW periods. Spectral attenuation reduced the MAEs during RW at every lead time, with statistically significant reductions at 24, 48, 96, and 120 h. During RI, TIFNet-IFSspec-ZS produced statistically significant larger errors at 24, 48, and 120 h (Fig. 7g,h). The effect of reducing short-wavelength variability was therefore small in terms of the aggregate but strongly dependent on storm regime, with the largest degradation for SuperTY and RI forecasts.

2.4.3 Effect of input-specific fine-tuning

We next tested whether the response to spectral attenuation depended on adaptation of TIFNet to the modified inputs. Fine-tuning the prediction head with IFSspec produced TIFNet-IFSspec-FT, which was compared with the corresponding zero-shot configuration, TIFNet-IFSspec-ZS, and with TIFNet-IFS-FT using the original IFS fields (Fig. 7). Across all samples, TIFNet-IFSspec-FT reduced the MAEs by 0.10–0.63 $m\,s^{-1}$ relative to TIFNet-IFSspec-ZS, with statistically significant reductions from 48 to 96 h (Fig. 7a). Among the intensity categories, the largest improvements were found for SuperTY. Fine-tuning reduced the MAEs by 0.71–3.31 $m\,s^{-1}$ at all evaluated lead times. TIFNet-IFSspec-FT also had lower MAEs than TIFNet-IFS-FT during 48–120 h, with reductions of 0.73–2.15 $m\,s^{-1}$ (Fig. 7f).

A similar response occurred during RI periods. TIFNet-IFSspec-FT reduced the MAEs by 1.37–3.71 $m\,s^{-1}$ relative to TIFNet-IFSspec-ZS at every evaluated lead time. TIFNet-IFSspec-FT also had lower MAEs than TIFNet-IFS-FT, with reductions of 0.53–4.77 $m\,s^{-1}$ (Fig. 7h). In contrast, fine-tuning did not consistently reduce errors for TS, STS, TY, or RW periods. The response to spectral attenuation was not determined by the atmospheric fields alone: adaptation to the modified input distribution recovered much of the degradation for SuperTY and RI and, at several lead times, produced lower errors than fine-tuning with the original IFS fields.

Overall, the sensitivity experiments separate three properties of the atmospheric inputs. Increasing forecast error degraded TIFNet intensity skill, whereas reducing short-wavelength variability had little effect on overall performance but increased errors for SuperTY and RI periods. Input-specific fine-tuning recovered much of this degradation. These results are consistent with the longer-range TIFNet-AIFS advantage arising from a trade-off: the lower atmospheric-field errors of AIFS offset part of the loss associated with smoother, short-wavelength-deficient fields, while adaptation to the AIFS inputs provides additional improvement.

3 Discussion

Recent AI weather models have demonstrated competitive skill in global atmospheric prediction and TC track forecasting (Bi et al., 2023; Lam et al., 2023; Price et al., 2025; Lang et al., 2024), but whether they can provide useful intensity forecasts despite smoother storm representations remains unclear. Here, we compared AIFS and IFS both as direct TC prediction systems and as atmospheric inputs to TIFNet. A central result of this study is that direct TC intensity skill and the usefulness of atmospheric forecast fields for intensity prediction are not equivalent. AIFS produced larger direct intensity errors than IFS, yet TIFNet driven by AIFS had lower errors at longer lead times, with the largest differences for SuperTY and pooled rapid-intensity-change samples (RI and RW). The reversal in model ranking shows that a global forecasting system need not provide the most accurate direct intensity forecast to supply atmospheric information that is useful for subsequent intensity prediction.

Atmospheric forecast accuracy provides one explanation for this behavior. AIFS had lower atmospheric-field errors than IFS across the forecast period, and the difference generally increased with lead time. The forecast-lead sensitivity experiments showed corresponding deterioration in TIFNet intensity forecasts as errors in the IFS input fields increased. This supports the view that accurate prediction of the evolving atmospheric environment is important for longer-range intensity prediction, consistent with the established role of environmental conditions in regulating TC evolution (Emanuel, 1999; Kaplan et al., 2015; Montgomery and Smith, 2017). It also helps explain why the advantage of TIFNet-AIFS emerged mainly at longer lead times, even though AIFS remained less accurate than IFS in direct intensity prediction.

The spatial characteristics of the input fields also affected intensity prediction. AI weather models often produce smoother forecasts and reduce small-scale variability compared with numerical prediction systems (Bi et al., 2023; Bonavita, 2024). However, reducing short-wavelength power in IFS did not produce a uniform response. Instead, its impact depended on storm intensity and intensity evolution: spectral attenuation improved forecasts for weaker storms and RW but degraded forecasts for SuperTY and RI. These results show that small-scale information remains important for intense and rapidly evolving TCs, even though accurate representation of environmental evolution also provides valuable predictive information.

How the intensity model is adapted to those fields also matters. Fine-tuning TIFNet on IFSspec recovered much of the degradation found in the zero-shot forecasts for SuperTY and RI and, at several lead times, produced lower errors than fine-tuning with the original IFS fields. The usefulness of an atmospheric forecast is thus determined not only by the information contained in the fields, but also by how effectively the intensity model extracts that information. This result also suggests that smoother atmospheric fields should not be interpreted by themselves as a direct measure of their value for TC intensity prediction. Reduced short-wavelength variability can remove information that is important for intense and rapidly evolving storms, but part of this loss can be compensated by adapting the prediction model to the characteristics of the input fields.

These findings point to a different way of using global atmospheric forecasts for TC intensity prediction. Conventional dynamical forecasting has sought to improve intensity prediction through increasingly realistic representations of the storm, supported by higher resolution, improved initialization and data assimilation, and advances in physical parameterizations (Gopalakrishnan et al., 2012; Goldenberg et al., 2015; Emanuel and Zhang, 2016). Global AI weather models are primarily optimized to predict the evolution of the atmospheric state and do not necessarily reproduce the same degree of storm-scale detail (Bi et al., 2023; Lam et al., 2023). TIFNet uses the resulting forecast information: rather than requiring the atmospheric model itself to reproduce the correct TC intensity, it learns a direct relationship between the predicted atmospheric environment, recent intensity evolution, and subsequent intensity change. More recent cyclone-specific AI forecasting systems have likewise shown that substantial information about TC evolution can be extracted from relatively coarse global atmospheric fields (Alet et al., 2026). Our results suggest that the value of a global weather forecast for TC intensity prediction lies not only in how accurately it represents the storm, but also in the intensity information retained in the predicted environment and how effectively that information can be extracted by a specialized model.

Several directions remain for future research. This study focused on maximum wind intensity prediction, and future work should investigate whether the relationship between atmospheric-field accuracy, spatial information, and model adaptation also applies to other aspects of TC prediction, including wind structure, precipitation, and landfall impacts. Further controlled experiments will also be needed to identify the atmospheric processes and spatial scales that contribute most to intensity predictability. More broadly, our results suggest that future forecasting systems might benefit from combining accurate global AI forecasts with specialized prediction models, allowing global atmospheric evolution and cyclone-specific intensity information to be represented within

complementary frameworks.

4 Methods

4.1 Datasets

The positions and intensities of TCs were obtained from the CMA best-track dataset for the western North Pacific, including the South China Sea (Ying et al., 2014; Lu et al., 2021). Six-hourly records at 00, 06, 12, and 18 UTC were used. For each TC, records were retained from the first to the last tropical-storm-or-stronger observation, including intervening tropical-depression records followed by reintensification. Subsequent extratropical records were excluded. The independent 2023–2024 test period contained 43 TCs and 906 retained CMA records.

Atmospheric data were obtained from the 0.25° hourly ERA5 (Hersbach et al., 2020), ECMWF IFS deterministic forecasts and analyses (ECMWF, 2023), and retrospective forecasts produced with ECMWF deterministic AIFS Single v1.0 from IFS analyses (Lang et al., 2024; Pinnington et al., 2025). ERA5 data from 1990–2020 were used for TIFNet pretraining and data from 2021–2022 used for validation. Input-specific fine-tuning used forecasts from 2020–2021 for training and data from 2022 for model selection. Forecasts from 2023–2024 were reserved exclusively for independent evaluation and were not used for training or model selection.

4.2 TC-centered atmospheric inputs

For each data source, atmospheric fields were independently centered on the diagnosed TC position before being supplied to TIFNet. TC centers were diagnosed separately from ERA5, IFS, AIFS, and verifying IFS analyses using three equally weighted estimates: the geometric circulation center from 850-hPa winds, the minimum sea level pressure, and the minimum 850-hPa geopotential height. The CMA position was used only to initialize the search at forecast issue time. At subsequent lead times, the search was initialized from the previously diagnosed center and refined within a 300-km radius.

The centered fields were linearly interpolated onto a 72 × 72 latitude–longitude grid at 0.25° spacing, covering approximately 18° × 18°. Each time step contained 44 atmospheric fields. Nine surface fields were used: 10-m zonal and meridional wind, 2-m air temperature, skin temperature, mean sea level pressure, surface geopotential, land–sea mask, subgrid orographic standard deviation, and orographic slope. Zonal wind, meridional wind, temperature, specific humidity, and geopotential at 100, 200, 300, 500, 700, 850 and 925 hPa provided a further 35 fields.

All inputs were normalized using statistics calculated from the 1990–2020 ERA5 pretraining dataset. The same statistics were applied to all data sources and forecast lead times. Wind, temperature, pressure, and upper-air variables were standardized using the ERA5 mean and standard deviation, whereas surface geopotential, land–sea mask, and the two orographic variables were scaled using the corresponding minimum and range.

4.3 TIFNet prediction and input-specific fine-tuning

We used the TIFNet architecture described by Qu et al. (2026), retaining its encoder–decoder structure and noniterative prediction of TC intensity change. The encoder received five atmospheric states at 6-h intervals together with observed CMA intensities from -24 to 0 h. The decoder received 21 forecast atmospheric states from 0 to 120 h at 6-h intervals, without future observed intensity. Spatial patches of surface and upper-air fields were processed using separate spatial and temporal attention, followed by cross-attention between the historical and forecast representations and a final linear regression head.

A single forward pass predicted TC intensity change at all forecast lead times. The target intensity change was defined as

$$\Delta V(t) = V(t) - V(0)$$

and the predicted intensity as

$$\widehat{V}(t) = V(0) + \widehat{\Delta V}(t)$$

where $V(0)$ is the observed intensity at forecast initialization. Forecasts at later lead times did not use predicted intensities from earlier lead times.

All experiments started from the same ERA5-pretrained checkpoint selected using the 2021–2022 validation loss. For zero-shot prediction, this checkpoint was applied directly to each atmospheric forecast source without fine-tuning. For input-specific fine-tuning, the atmospheric feature extractors, encoder, decoder, and cross-attention layers were frozen, and only the final linear regression head was optimized separately for IFS, AIFS, and IFSspec using the corresponding 2020–2021 inputs. Data from 2022 were used for model selection.

For both pretraining and input-specific fine-tuning, the loss function was the mean squared error of the standardized intensity changes over valid forecast lead times. Optimization was performed using the Adam optimizer (Kingma and Ba, 2015) with batch size of 8 and weight decay of 10^{-3}. The learning rate increased from 10^{-6} to 10^{-4} over the first three epochs and subsequently followed cosine annealing (Loshchilov and Hutter, 2017). Training was terminated after 10 consecutive epochs without improvement in validation loss. The checkpoint with the lowest validation loss was retained.

4.4 Experimental design

To identify the factors underlying the differences between TIFNet-AIFS and TIFNet-IFS, we designed three controlled experiments (Table 1). First, AIFS and IFS were compared both as direct TC forecasting systems and as atmospheric inputs to TIFNet. Second, IFS fields from different forecast lead times were used to examine the sensitivity of TIFNet to input forecast-error growth. Third, original and spectrally modified IFS fields (IFSspec) were compared under zero-shot and fine-tuned settings to assess the effects of spectral information and model adaptation.

4.4.1 AIFS and IFS comparison

Direct forecasts of TC track and intensity from AIFS and IFS were evaluated against CMA best-track observations. The corresponding TC-centered atmospheric forecasts were then used independently as inputs to TIFNet to produce TIFNet-AIFS and TIFNet-IFS forecasts. The TIFNet architecture, prediction target, and evaluation samples were identical between the two configurations.

Intensity error at forecast lead t was defined as

$$E_V(t) = \left| \widehat{V}(t) - V_{\mathrm{obs}}(t) \right|$$

where $\widehat{V}(t)$ and $V_{\mathrm{obs}}(t)$ are the forecast and observed TC intensities, respectively. Track error was calculated as the great-circle distance between the diagnosed forecast center and the corresponding CMA best-track position. Errors were summarized at each forecast lead by their mean values.

Forecasts were stratified by CMA intensity category at verification time. The intensity ranges were TS, 17.2–24.5 m s^{-1}; STS, 24.5–32.7 m s^{-1}; TY, 32.7–41.5 m s^{-1}; STY, 41.5–51.0 m s^{-1}; and SuperTY, $\geq$51.0 m s^{-1} (Lu et al., 2021). The observed 24-h intensity change ending at verification time t was defined as

$$\Delta V_{\mathrm{obs,24h}}(t) = V_{\mathrm{obs}}(t) - V_{\mathrm{obs}}(t - 24\,\mathrm{h}).$$

Periods of RI and RW were defined by $\Delta V_{\mathrm{obs,24h}}(t) \geq 15\,\mathrm{m\,s^{-1}}$ and $\Delta V_{\mathrm{obs,24h}}(t) \leq -15\,\mathrm{m\,s^{-1}}$, respectively; thus, rapid-intensity-change samples satisfied $|\Delta V_{\mathrm{obs,24h}}(t)| \geq 15\,\mathrm{m\,s^{-1}}$. Here, t denotes the verification time. The threshold and sample-selection procedure were otherwise unchanged.

AIFS and IFS atmospheric forecasts were compared with verifying IFS analyses on the common TC-centered grid. Forecast and analysis fields were independently centered on their diagnosed TC positions before verification, excluding absolute TC position differences from the field-error calculation.

For each variable and forecast lead, the RMSE was calculated by pooling squared differences over all retained samples and all 72×72 grid points. The analysis included 16 representative fields: 2-m temperature, mean sea level pressure, and 10-m zonal and meridional winds, together with temperature, geopotential, and zonal and meridional winds at 850, 500, and 200 hPa.

4.4.2 Input forecast-lead sensitivity experiment

A forecast-lead sensitivity experiment was performed using IFS atmospheric fields to examine the response of TIFNet to increasing atmospheric forecast error. The same zero-shot TIFNet checkpoint was used in all experiments. Decoder inputs were constructed using IFS fields at fixed input-field lead times of 24, 48, 72, or 96 h, defining TIFNet-IFS-24h, TIFNet-IFS-48h, TIFNet-IFS-72h, and TIFNet-IFS-96h, respectively. The encoder inputs and observed initial intensities were kept identical, and no TIFNet parameters were updated in these experiments.

4.4.3 Input spectral sensitivity experiment

To examine the effect of spatial-scale variability in the atmospheric inputs, we constructed IFSspec, a spectrally modified version of IFS. Paired 850-hPa kinetic energy spectra showed lower AIFS spectral energy than IFS at smaller spatial scales (Fig. A1). For each paired dynamic field, the Fourier amplitudes of IFS were adjusted as a function of radial wavenumber toward those of AIFS at the same valid time, variable, and pressure level, while retaining the spatial phase of IFS.

For each field, the spatial mean was removed and a two-dimensional Hann-windowed periodogram was calculated. Spectral power was radially aggregated into 30 wavenumber bins using an approximate horizontal grid spacing of 25 km. The scale-dependent amplitude gain was defined as

$$G(k) = \min\{1, [\frac{E_{\mathrm{AIFS}}(k)}{\max(E_{\mathrm{IFS}}(k), 10^{-12})}]^{1/2}\},$$

where $E_{\mathrm{IFS}}(k)$ and $E_{\mathrm{AIFS}}(k)$ are the radial power spectra of the paired IFS and AIFS fields, respectively, at radial wavenumber k.

The gain was Gaussian-smoothed over two wavenumber bins, clipped to [0, 1], and interpolated onto the corresponding two-dimensional Fourier frequencies before being applied to the IFS Fourier amplitudes. Because $G(k) \leq 1$, the transformation attenuated scales only for which IFS contained more spectral energy than AIFS and did not amplify scales for which IFS energy was lower. Fourier phases were unchanged, and the original IFS spatial mean was restored after inverse transformation.

The diagnosed TC center, 0.25° grid, surface geopotential, land–sea mask, and orographic fields were retained unchanged. The perturbation modified spectral amplitudes while retaining the TC position, Fourier phase, and static surface fields. The 850-hPa kinetic energy spectra of IFSspec were recalculated using the same paired samples to verify that the transformation shifted the smaller-scale spectral characteristics of IFS towards those of AIFS (Fig. A1).

IFSspec was used only as a diagnostic sensitivity experiment. The zero-shot configurations TIFNet-IFS-ZS and TIFNet-IFSspec-ZS isolated the effect of spectral perturbation, whereas TIFNet-IFS-FT and TIFNet-IFSspec-FT assessed the effect of spectral perturbation after input-specific fine-tuning.

Table 1. Experimental configurations used to evaluate the effects of atmospheric forecast source, input forecast lead, spectral characteristics, and model adaptation on TIFNet intensity prediction.

Experiment	**Configuration**	**Atmospheric input**	**TIFNet setting**	**Controlled factor(s)**
Atmospheric-source comparison	TIFNet-IFS	IFS forecasts	Head fine-tuned	Atmospheric input source
Atmospheric-source comparison	TIFNet-AIFS	AIFS forecasts	Head fine-tuned	Atmospheric input source
Input forecast-lead sensitivity	TIFNet-IFS-24h	IFS fields at 24-h forecast lead	Zero-shot	IFS input forecast lead
Input forecast-lead sensitivity	TIFNet-IFS-48h	IFS fields at 48-h forecast lead	Zero-shot	IFS input forecast lead
Input forecast-lead sensitivity	TIFNet-IFS-72h	IFS fields at 72-h forecast lead	Zero-shot	IFS input forecast lead
Input forecast-lead sensitivity	TIFNet-IFS-96h	IFS fields at 96-h forecast lead	Zero-shot	IFS input forecast lead
Spectral perturbation × adaptation	TIFNet-IFS-ZS	Original IFS fields	Zero-shot	Spectral content × model adaptation
Spectral perturbation × adaptation	TIFNet-IFSspec-ZS	IFSspec: IFS fields with Fourier amplitudes adjusted toward AIFS	Zero-shot	Spectral content × model adaptation
Spectral perturbation × adaptation	TIFNet-IFS-FT	Original IFS fields	Head fine-tuned	Spectral content × model adaptation
Spectral perturbation × adaptation	TIFNet-IFSspec-FT	IFSspec: IFS fields with Fourier amplitudes adjusted toward AIFS	Head fine-tuned	Spectral content × model adaptation

Note: IFSspec denotes spectrally modified IFS fields in which the Fourier amplitudes are adjusted toward those of AIFS while retaining the IFS spatial mean, Fourier phase, grid, and static fields. ZS denotes zero-shot application of the ERA5-pretrained TIFNet without fine-tuning. FT denotes input-specific fine-tuning in which only the final regression head is optimized while all other TIFNet parameters remain frozen.

4.5 Statistical analysis

Uncertainty in the reported mean errors was estimated using 20,000 paired record-level bootstrap resamples drawn with replacement. The same resampled indices were used for forecasts in each paired comparison. The 95% confidence interval was defined by the 2.5th and 97.5th percentiles of the bootstrap distribution. A random seed was used for all resampling.

Differences between paired forecasts were tested using two-sided paired-bootstrap tests of the mean absolute error difference. The null distribution was obtained by centering the observed paired error differences at zero. The finite-resample correction

$$P = \frac{b+1}{B+1}$$

was applied, where $B = 20{,}000$ is the number of bootstrap replicates and b is the number of replicates at least as extreme as the observed difference.

The Benjamini–Hochberg false-discovery-rate correction (Benjamini and Hochberg, 1995) was applied separately within each plotted panel across the displayed forecast lead times and prespecified comparisons. Adjusted $P_{\mathrm{adj}} < 0.05$ was considered statistically significant. The two single-storm case studies (Figs. 3 and 4) were presented descriptively, without bootstrap confidence intervals, hypothesis tests or multiple-comparison correction; atmospheric-field RMSEs were also treated as descriptive.

Data availability

CMA best-track data are available from the CMA Tropical Cyclone Data Center at `https://tcdata.typhoon.org.cn/zjljsjj_zlhq.html`, ERA5 from the Copernicus Climate Data Store at `https://cds.climate.copernicus.eu/`, and IFS analyses and forecasts from the ECMWF archive at `https://www.ecmwf.int/en/forecasts/dataset/operational-archive`, subject to ECMWF access conditions. The AIFS Single v1.0 model and configuration are available at `https://huggingface.co/ecmwf/aifs-single-1.0`.

Code availability

The IFSMatch2AIFS spectral-matching code, TIFNet model pseudocode, trained ONNX models, and inference example are available at `https://github.com/Alley0911/TIFNet-AIFS-VS-TIFNet-IFS`.

Acknowledgments

This study was supported by the National Key R&D Program of China under grant 2023YFC3008005, the Joint Research Project for Meteorological Capacity Improvement under grant 2025NLTSQ005, the National Natural Science Foundation of China under grants 42375015, and 42192554, the Typhoon Scientific and Technological Innovation Group of the China Meteorological Administration under grant CMA2023ZD06, the Basic Research Fund of CAMS under grant 2023Z020, S&T Development Fund of CAMS under Grant 2024KJ018 and 2024KJ022, and State Key Laboratory of Severe Weather Meteorological Science and Technology 2026QZA03, and 2026QZA20.

Author contributions

H.Q. and H.X. conceived and designed the study. H.Q. performed the data processing, model experiments, statistical analyses and visualization. H.Q. and H.X. interpreted the results and drafted the manuscript. L.D., C.X. and D.Z. contributed to interpretation of the results and review and editing of the manuscript. All authors discussed the results and approved the final manuscript.

Competing interests

The authors declare no competing interests.

References

Ferran Alet, Tom R. Andersson, Ilan Price, Stratis Markou, Andrew El-Kadi, Dominic Masters, Amy Li, Samier Merchant, Natalie Williams, Gregory Thornton, Ken MacKay, Olivia Graham,

Akib Uddin, Ben Gaiarin, Devaja Shah, Elinor Kruse, Wallace Hogsett, David Zelinsky, John Cangialosi, Jonathan Martinez, James Franklin, Mark DeMaria, Kate Musgrave, Caroline L. Bain, Helen Titley, Jacklynn Stott, Remi Lam, Aaron Bell, Paul Komarek, Matthew Willson, Alvaro Sanchez-Gonzalez, and Peter Battaglia. Operational tropical cyclone forecasting with AI. *Nature*, 657(8132):680–688, 2026. doi: 10.1038/s41586-026-10953-2.

Peter Bauer, Alan Thorpe, and Gilbert Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55, 2015. doi: 10.1038/nature14956.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619:533–538, 2023. doi: 10.1038/s41586-023-06185-3.

Massimo Bonavita. On some limitations of current machine learning weather prediction models. *Geophysical Research Letters*, 51(12):e2023GL107377, 2024. doi: 10.1029/2023GL107377.

John P. Cangialosi, Eric Blake, Mark DeMaria, Andrew Penny, Andrew Latto, Edward Rappaport, and Vijay Tallapragada. Recent progress in tropical cyclone intensity forecasting at the national hurricane center. *Weather and Forecasting*, 35(5):1913–1922, 2020. doi: 10.1175/WAF-D-20-0059.1.

Mark DeMaria. The effect of vertical shear on tropical cyclone intensity change. *Journal of the Atmospheric Sciences*, 53(14):2076–2088, 1996. doi: 10.1175/1520-0469(1996)053⟨2076:TEOVSO⟩2.0.CO;2.

Mark DeMaria, Michelle Mainelli, Lynn K. Shay, John A. Knaff, and John Kaplan. Further improvements to the statistical hurricane intensity prediction scheme (SHIPS). *Weather and Forecasting*, 20(4):531–543, 2005. doi: 10.1175/WAF862.1.

Mark DeMaria, Charles R. Sampson, John A. Knaff, and Kate D. Musgrave. Is tropical cyclone intensity guidance improving? *Bulletin of the American Meteorological Society*, 95(3):387–398, 2014. doi: 10.1175/BAMS-D-12-00240.1.

Mark DeMaria, James L. Franklin, Galina Chirokova, Jacob Radford, Robert DeMaria, Kate D. Musgrave, and Imme Ebert-Uphoff. An operations-based evaluation of tropical cyclone track and intensity forecasts from artificial intelligence weather prediction models. *Artificial Intelligence for the Earth Systems*, 4(4):240085, 2025. doi: 10.1175/AIES-D-24-0085.1.

ECMWF. IFS documentation CY48R1 – part III: Dynamics and numerical procedures. In *IFS Documentation CY48R1*. ECMWF, 2023. doi: 10.21957/26f0ad3473.

Kerry Emanuel and Fuqing Zhang. On the predictability and error sources of tropical cyclone intensity forecasts. *Journal of the Atmospheric Sciences*, 73(9):3739–3747, 2016. doi: 10.1175/JAS-D-16-0100.1.

Kerry A. Emanuel. Thermodynamic control of hurricane intensity. *Nature*, 401(6754):665–669, 1999. doi: 10.1038/44326.

Stanley B. Goldenberg, Sundararaman G. Gopalakrishnan, Vijay Tallapragada, Thiago Quirino, Frank Marks, Samuel Trahan, Xuejin Zhang, and Robert Atlas. The 2012 triply nested, high-resolution operational version of the hurricane weather research and forecasting model (HWRF): Track and intensity forecast verifications. *Weather and Forecasting*, 30(3):710–729, 2015. doi: 10.1175/WAF-D-14-00098.1.

Sundararaman G. Gopalakrishnan, Stanley Goldenberg, Thiago Quirino, Xuejin Zhang, Frank D. Marks, Jr., Kao-San Yeh, Robert Atlas, and Vijay Tallapragada. Toward improving high-resolution numerical hurricane forecasting: Influence of model horizontal grid resolution, initialization, and physics. *Weather and Forecasting*, 27(3):647–666, 2012. doi: 10.1175/WAF-D-11-00055.1.

Hans Hersbach, Bill Bell, Paul Berrisford, et al. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: 10.1002/qj.3803.

Cheng Huang, Pan Mu, Jinglin Zhang, Sixian Chan, Shiqi Zhang, Hanting Yan, Shengyong Chen, and Cong Bai. Benchmark dataset and deep learning method for global tropical cyclone forecasting. *Nature Communications*, 16(1):5923, 2025. doi: 10.1038/s41467-025-61087-4.

Haiyan Jiang. The relationship between tropical cyclone intensity change and the strength of inner-core convection. *Monthly Weather Review*, 140(4):1164–1176, 2012. doi: 10.1175/MWR-D-11-00134.1.

John Kaplan, Mark DeMaria, and John A. Knaff. A revised tropical cyclone rapid intensification index for the atlantic and eastern north pacific basins. *Weather and Forecasting*, 25(1):220–241, 2010. doi: 10.1175/2009WAF2222280.1.

John Kaplan, Christopher M. Rozoff, Mark DeMaria, Charles R. Sampson, James P. Kossin, Christopher S. Velden, Joseph J. Cione, Jason P. Dunion, John A. Knaff, Jun A. Zhang, John F. Dostalek, Jeffrey D. Hawkins, Thomas F. Lee, and Jeremy E. Solbrig. Evaluating environmental impacts on tropical cyclone rapid intensification predictability utilizing statistical models. *Weather and Forecasting*, 30(5):1374–1396, 2015. doi: 10.1175/WAF-D-15-0032.1.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL `https://arxiv.org/abs/1412.6980`.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023. doi: 10.1126/science.adi2336.

Simon Lang, Mihai Alexe, Matthew Chantry, Jesper Søren Dramsch, Florian Pinault, Baudouin Raoult, Mariana C. A. Clare, Christian Lessig, Michael Maier-Gerber, Linus Magnusson, Zied Ben Bouallègue, Ana Prieto Nemesio, Peter D. Dueben, Andy Brown, Florian Pappenberger, and Florence Rabier. AIFS – ECMWF's data-driven forecasting system. *arXiv preprint arXiv:2406.01465*, 2024. doi: 10.48550/arXiv.2406.01465.

Juhyun Lee, Jungho Im, and Yeji Shin. Enhancing tropical cyclone intensity forecasting with explainable deep learning integrating satellite observations and numerical model outputs. *iScience*, 27(6):109905, 2024. doi: 10.1016/j.isci.2024.109905.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. URL `https://arxiv.org/abs/1608.03983`.

X. Q. Lu, H. Yu, M. Ying, B. K. Zhao, S. Zhang, L. M. Lin, L. N. Bai, and R. J. Wan. Western north pacific tropical cyclone database created by the china meteorological administration. *Advances in Atmospheric Sciences*, 38(4):690–699, 2021. doi: 10.1007/s00376-020-0211-7.

Michael T. Montgomery and Roger K. Smith. Recent developments in the fluid dynamics of tropical cyclones. *Annual Review of Fluid Mechanics*, 49(1):541–574, 2017. doi: 10.1146/annurev-fluid-010816-060022.

Pascal Peduzzi, Bruno Chatenoux, H. Dao, A. De Bono, C. Herold, J. Kossin, F. Mouton, and O. Nordbeck. Global trends in tropical cyclone risk. *Nature Climate Change*, 2:289–294, 2012. doi: 10.1038/nclimate1410.

Ewan Pinnington, Ana Prieto Nemesio, and Gabriel Moldovan. Operational release of AIFS Single 1.0. *ECMWF Newsletter*, 183:6–7, 2025. URL `https://www.ecmwf.int/en/newsletter/183/news/operational-release-aifs-single-10`.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R. Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Remi Lam, and Matthew Willson. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044): 84–90, 2025. doi: 10.1038/s41586-024-08252-9.

Hongyu Qu, Hongxiong Xu, Lin Dong, Chunyi Xiang, and Gaozhen Nie. Accurate tropical cyclone intensity forecasts using a non-iterative spatiotemporal transformer model. *npj Climate and Atmospheric Science*, 9:9, 2026. doi: 10.1038/s41612-025-01279-3.

Vijay Tallapragada, Ligia Bernardet, Mrinal K. Biswas, Sundararaman Gopalakrishnan, Young Kwon, Qingfu Liu, Timothy Marchok, David Sheinin, Mingjing Tong, Sam Trahan, Robert Tuleya, Robert Yablonsky, and Xuejin Zhang. Hurricane weather research and forecasting (HWRF) model: 2014 scientific documentation. Technical Report HWRF v3.6a, NCAR Developmental Testbed Center, 2014.

Xinyu Wang, Lei Liu, Kang Chen, Tao Han, Bin Li, and Lei Bai. VQLTI: Long-term tropical cyclone intensity forecasting with physical constraints. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27):28476–28484, 2025. doi: 10.1609/aaai.v39i27.35070.

Wenwei Xu, Karthik Balaguru, Andrew August, Nicholas Lalo, Nathan O. Hodas, Mark DeMaria, and David R. Judi. Deep learning experiments for tropical cyclone intensity forecasts. *Weather and Forecasting*, 36(4):1453–1470, 2021. doi: 10.1175/WAF-D-20-0104.1.

Munehiko Yamaguchi, Yasutaka Ikuta, Kosuke Ito, and Masaki Satoh. Tropical cyclone track and intensity predictions in the western north pacific basin using Pangu-Weather and JMA initial conditions. *Journal of the Meteorological Society of Japan*, 103(3):357–370, 2025. doi: 10.2151/jmsj.2025-018.

M. Ying, W. Zhang, H. Yu, X. Lu, J. Feng, Y. Fan, Y. Zhu, and D. Chen. An overview of the china meteorological administration tropical cyclone database. *Journal of Atmospheric and Oceanic Technology*, 31:287–301, 2014. doi: 10.1175/JTECH-D-12-00119.1.

Qiang Zhang, Liguang Wu, and Qiufeng Liu. Tropical cyclone damages in china 1983–2006. *Bulletin of the American Meteorological Society*, 90(4):489–496, 2009. doi: 10.1175/2008BAMS2631.1.

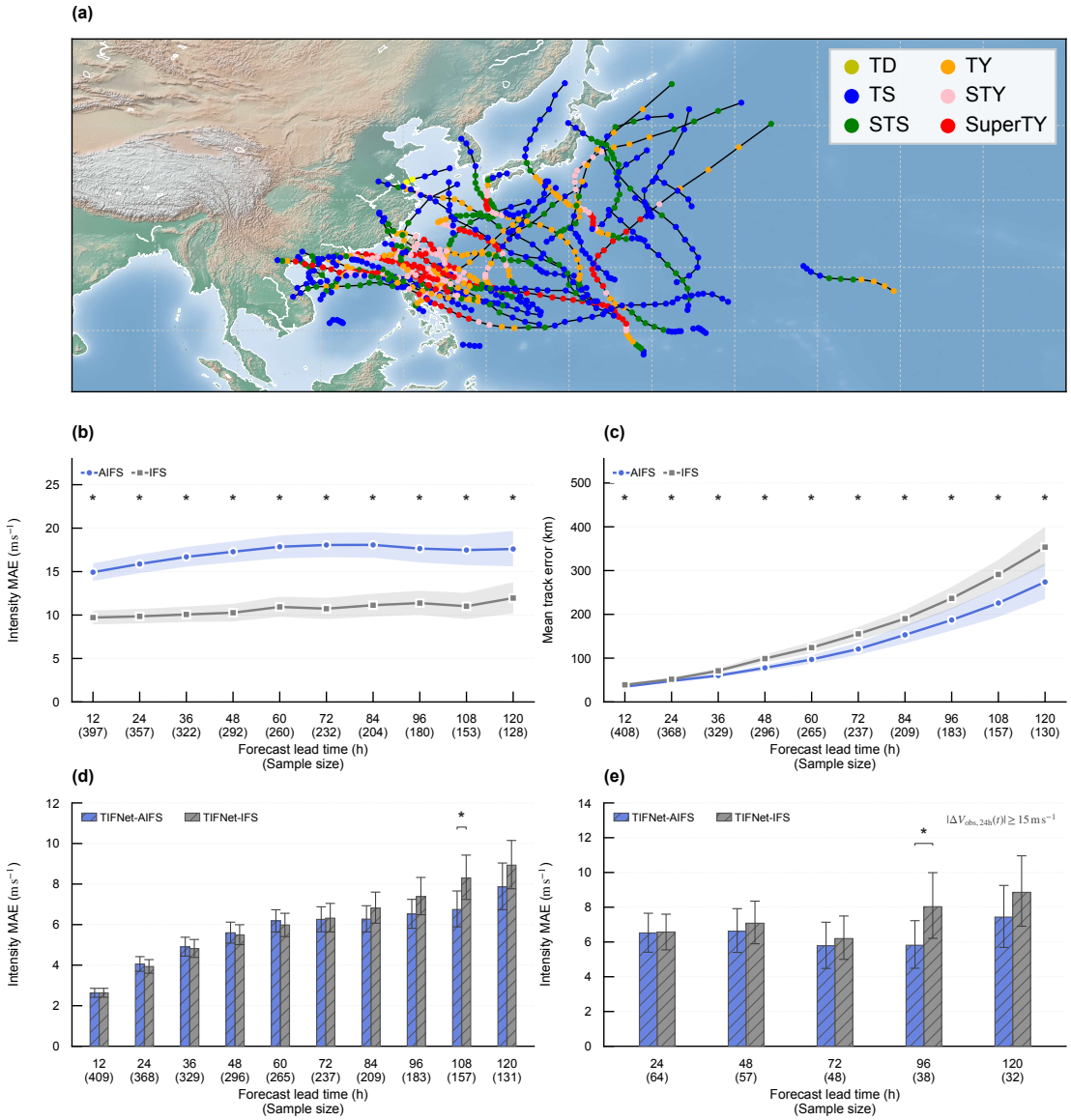


Figure 1. Overall comparison of AIFS and IFS forecasts and the corresponding TIFNet intensity forecasts. a, CMA best-track trajectories during 2023–2024, with fixes color-coded by intensity category. b, Intensity MAEs of the direct AIFS and IFS forecasts. c, Mean track errors of AIFS and IFS. d, Intensity MAEs of fine-tuned TIFNet driven by AIFS (TIFNet-AIFS) and IFS (TIFNet-IFS) for all samples. e, Intensity MAEs for rapid-intensity-change (RW or RI) samples. RI and RW samples are identified from the observed 24-h intensity change ending at each verification time, with intensity forecast errors evaluated at that time. Shaded bands in b,c and error bars in d,e indicate 95% percentile bootstrap confidence intervals. In d,e, brackets connect the compared configurations. Asterisks indicate statistically significant differences after correction for multiple comparisons ($P_{\mathrm{adj}} < 0.05$).

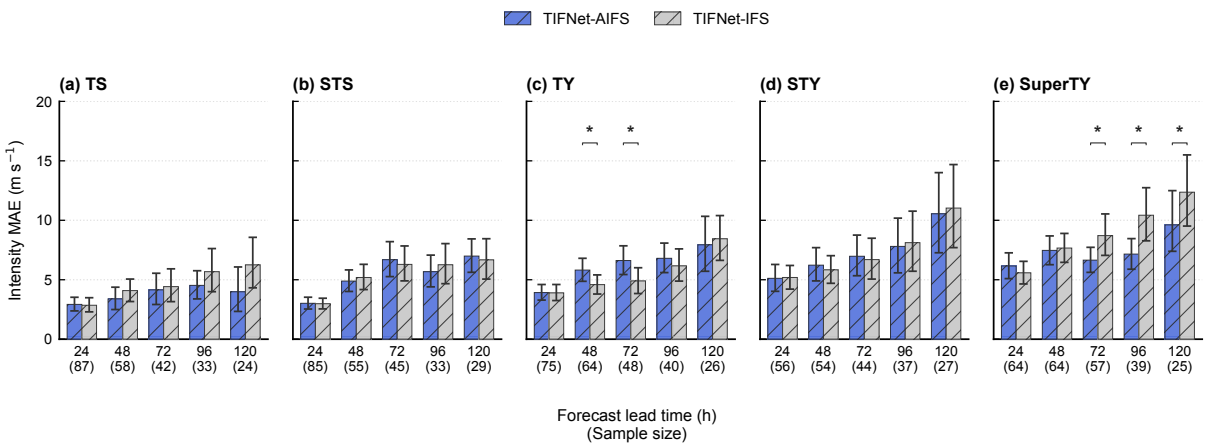


Figure 2. TIFNet intensity errors stratified by TC intensity category. Mean absolute intensity errors of fine-tuned TIFNet-AIFS and TIFNet-IFS forecasts are shown for a, TS; b, STS; c, TY; d, STY; and e, SuperTY, with categories defined from intensity at verification time. Blue bars on the left show TIFNet-AIFS and gray bars on the right show TIFNet-IFS; hatching denotes fine-tuned configurations. Error bars indicate 95% percentile bootstrap confidence intervals. Brackets connect the compared configurations, and asterisks indicate statistically significant differences after correction for multiple comparisons ($P_{\mathrm{adj}} < 0.05$).

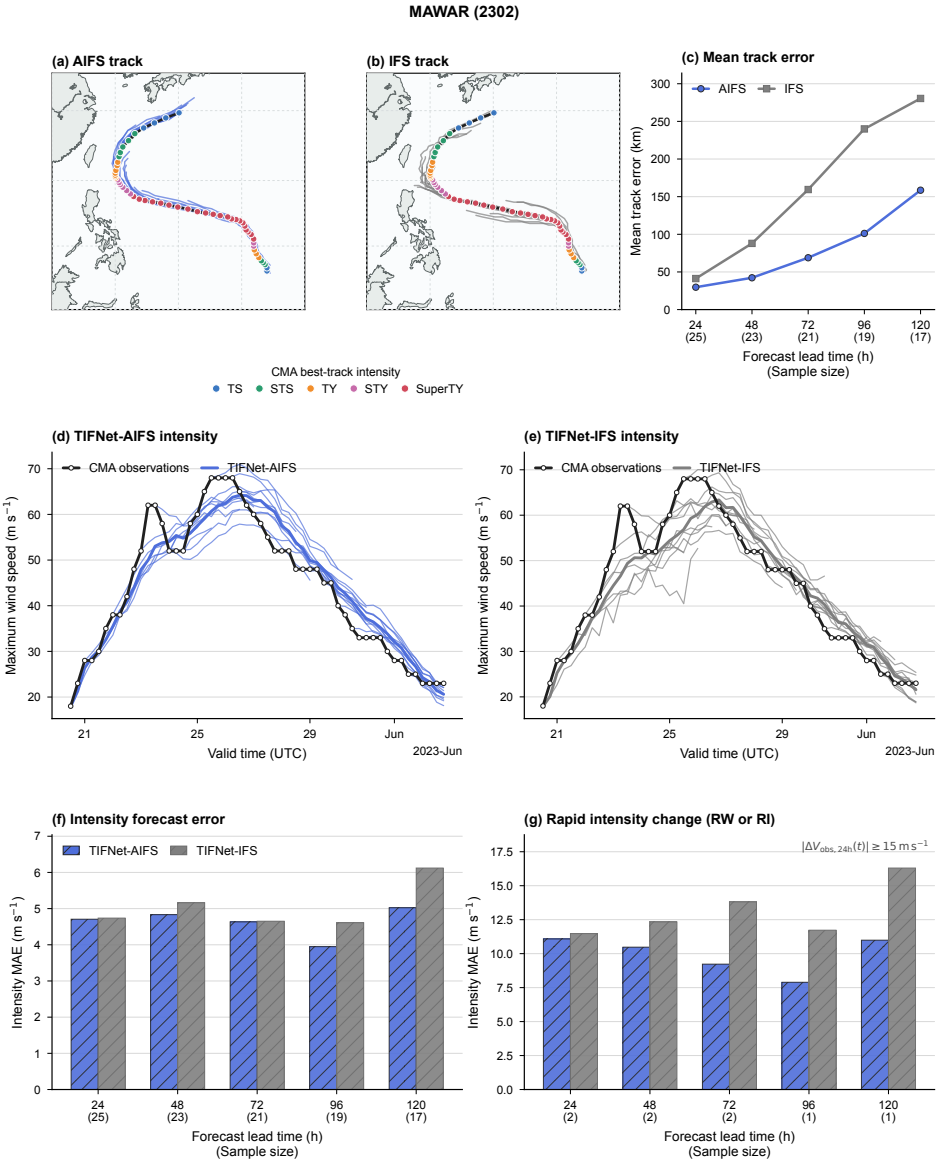


Figure 3. Case study of Typhoon Mawar (2023). a,b, AIFS and IFS forecast tracks, respectively. Thin colored lines show trajectories from individual initialization times; the black line and colored fixes show the CMA best track and its observed intensity categories. c, Mean track error, with AIFS shown by blue circles and IFS by gray squares. d,e, Fine-tuned TIFNet-AIFS and TIFNet-IFS intensity forecasts, respectively. Thin colored lines show individual forecasts, thick colored lines show the mean forecast at each valid time, and the black line with open circles shows CMA observations. f,g, Intensity MAE for all samples and rapid-intensity-change (RW or RI) samples, respectively. Blue AIFS bars are on the left and gray IFS bars are on the right; hatching denotes fine-tuned configurations. The case study is descriptive and no inferential tests are applied.

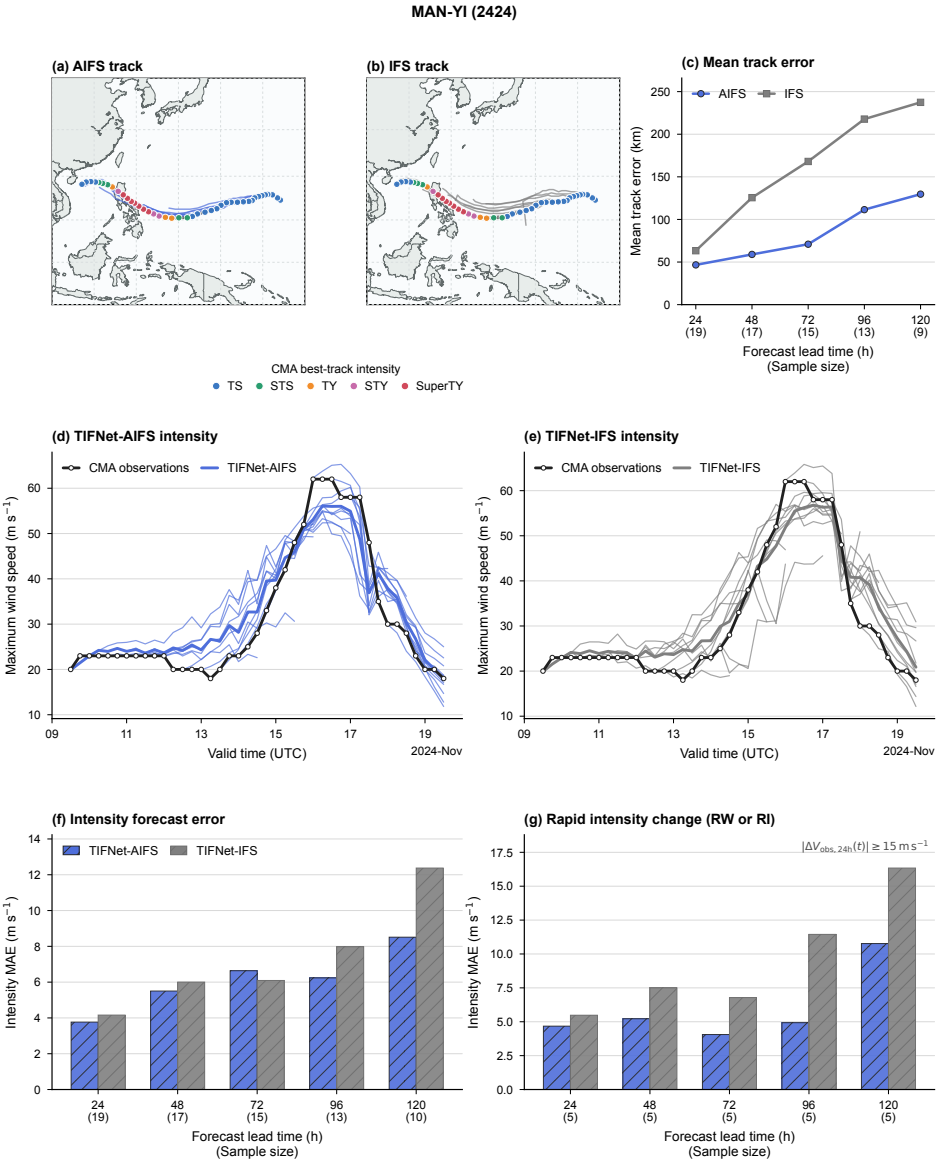


Figure 4. Case study of Typhoon Man-yi (2024). a,b, AIFS and IFS forecast tracks, respectively. Thin colored lines show trajectories from individual initialization times; the black line and colored fixes show the CMA best track and its observed intensity categories. c, Mean track error, with AIFS shown by blue circles and IFS by gray squares. d,e, Fine-tuned TIFNet-AIFS and TIFNet-IFS intensity forecasts, respectively. Thin colored lines show individual forecasts, thick colored lines show the mean forecast at each valid time, and the black line with open circles shows CMA observations. f,g, Intensity MAE for all samples and rapid-intensity-change (RW or RI) samples, respectively. Blue AIFS bars are on the left and gray IFS bars are on the right; hatching denotes fine-tuned configurations. The case study is descriptive and no inferential tests are applied.

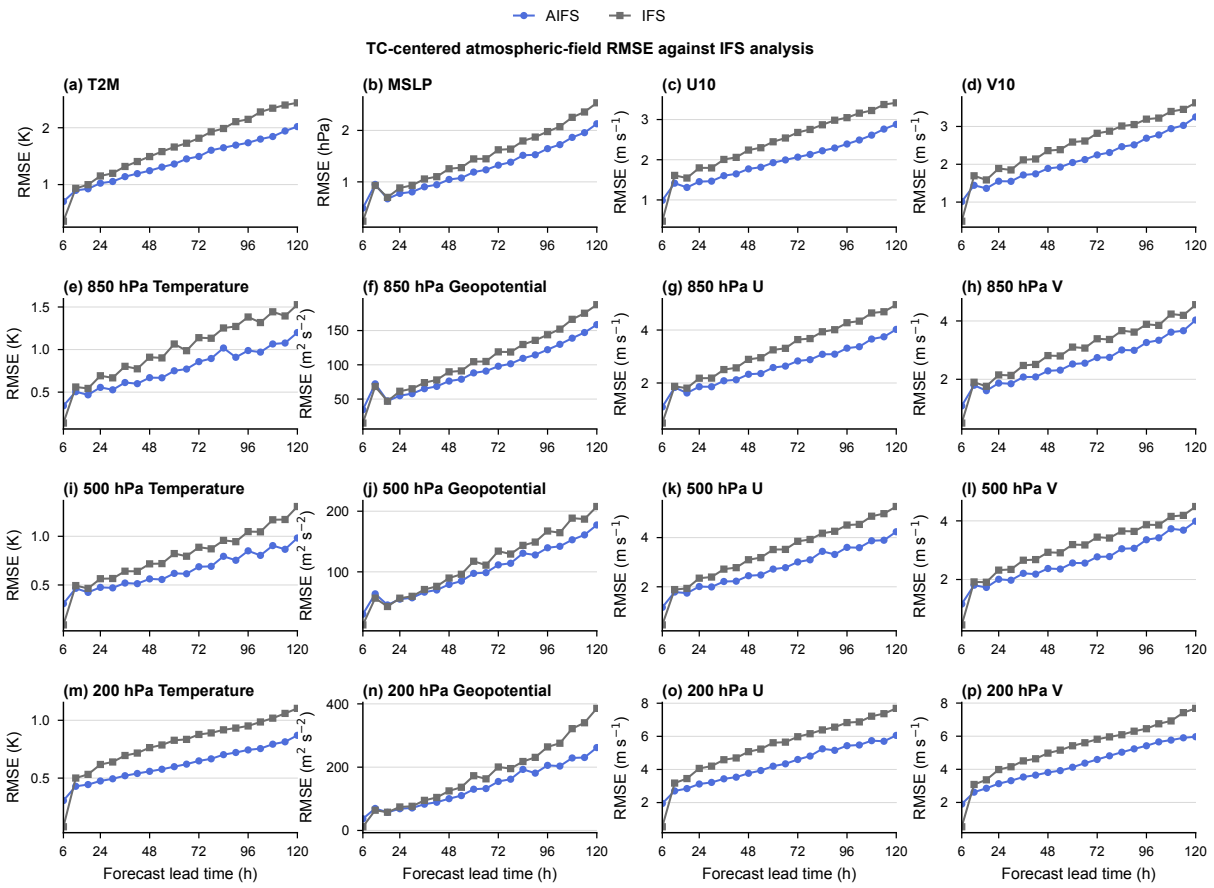


Figure 5. Errors in the AIFS and IFS atmospheric inputs to TIFNet. RMSEs relative to verifying IFS analyses are shown for a–d, 2-m temperature (T2M), mean sea level pressure (MSLP), 10-m zonal wind (U10), and 10-m meridional wind (V10), respectively; e–h, temperature, geopotential, zonal wind, and meridional wind at 850 hPa, respectively; i–l, the corresponding variables at 500 hPa; and m–p, the corresponding variables at 200 hPa. Curves show AIFS (blue circles) and IFS (gray squares) errors as a function of forecast lead time during the 2023–2024 evaluation period. Pressure RMSE is reported in hPa. All fields are verified on their respective TC-centered grids.

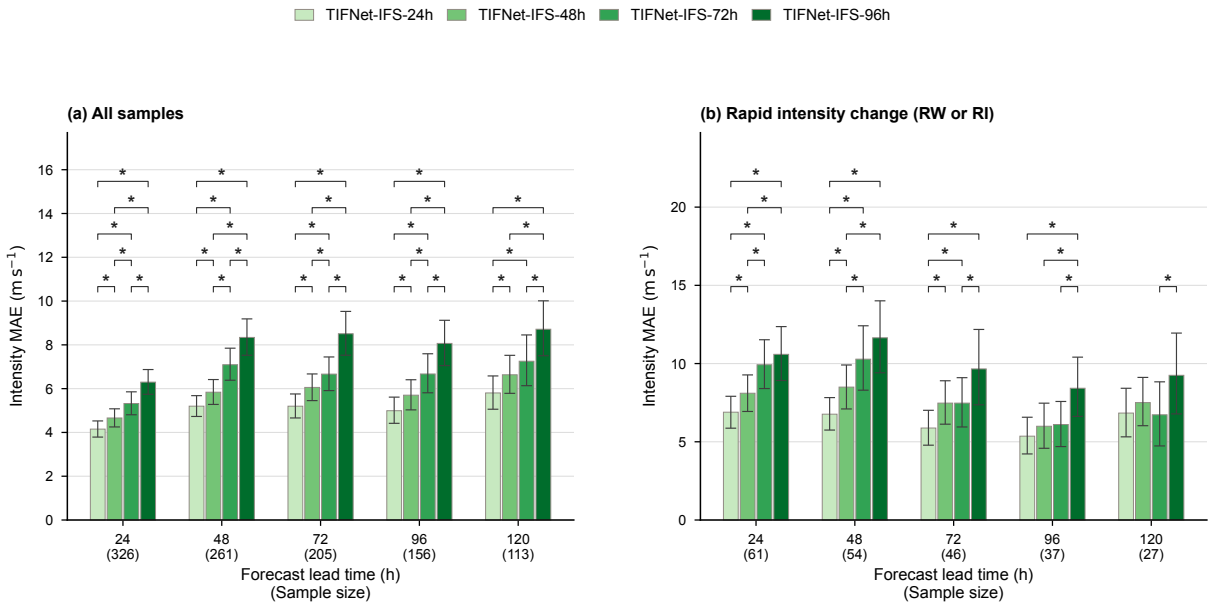


Figure 6. Sensitivity of TIFNet intensity forecasts to input forecast-error growth. Mean absolute intensity errors are shown for TIFNet-IFS-24h, TIFNet-IFS-48h, TIFNet-IFS-72h, and TIFNet-IFS-96h. The 24h, 48h, 72h and 96h suffixes in the model names denote the forecast lead times of the IFS input fields, whereas the horizontal axis denotes the TIFNet intensity forecast lead time. a, All forecast samples. b, Rapid-intensity-change (RW or RI) samples. Error bars indicate 95% percentile bootstrap confidence intervals. Brackets connect the configurations included in each of the six pairwise comparisons among the four input-lead configurations; asterisks indicate statistically significant differences after correction for multiple comparisons ($P_{\mathrm{adj}} < 0.05$).

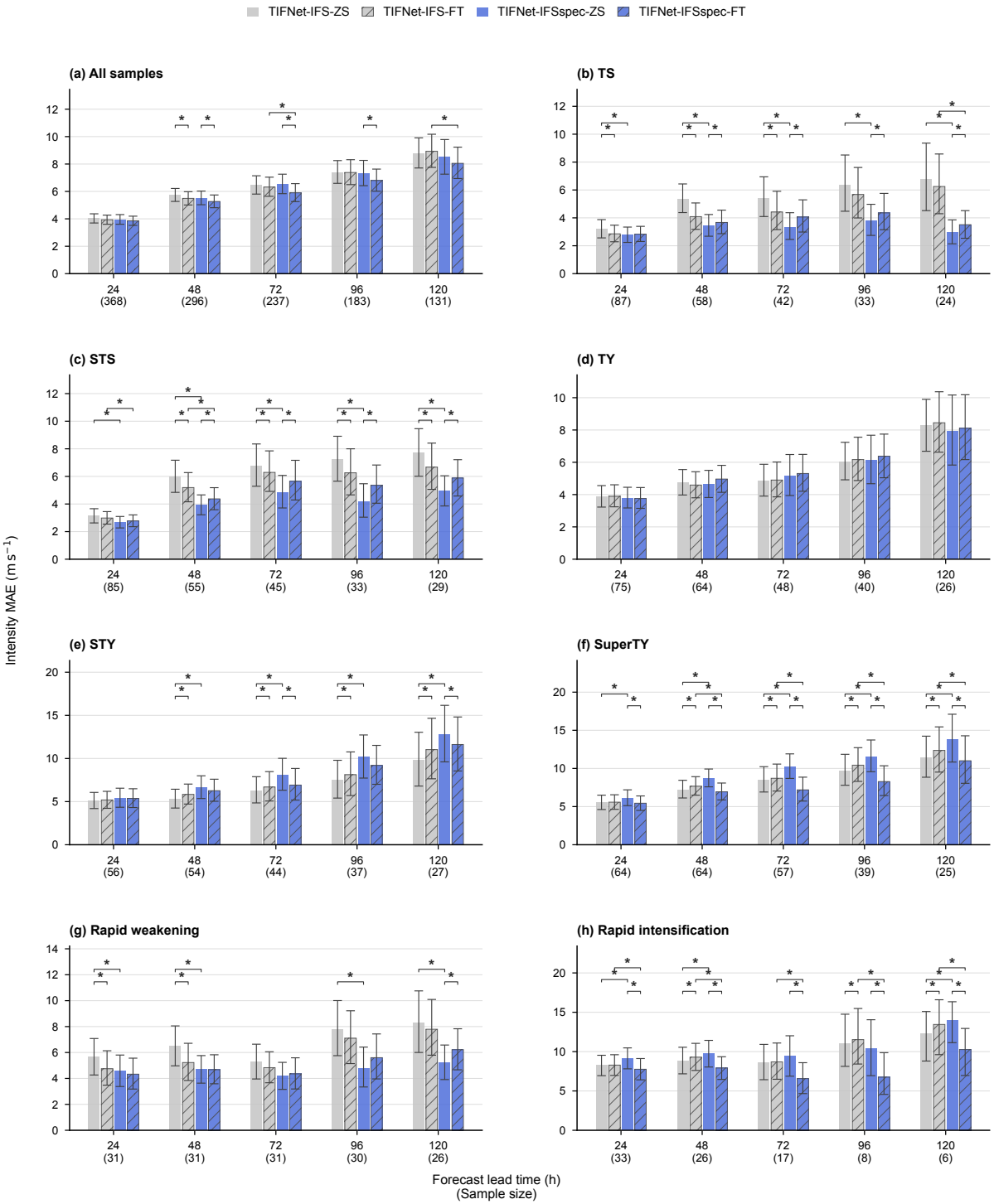


Figure 7. Sensitivity of TIFNet intensity forecasts to input spectra and fine-tuning. Mean absolute intensity errors are shown, from left to right within each lead-time group and in the legend, for TIFNet-IFS-ZS, TIFNet-IFS-FT, TIFNet-IFSspec-ZS, and TIFNet-IFSspec-FT. Gray denotes IFS inputs and blue denotes IFSspec inputs; solid bars denote zero-shot (ZS) application and hatched bars denote fine-tuning (FT). a, All forecast samples. b–f, Forecasts stratified by CMA intensity category at verification time: TS, STS, TY, STY, and SuperTY, respectively. g, RW samples. h, RI samples. Error bars indicate 95% percentile bootstrap confidence intervals. Brackets connect the compared configurations: IFS and IFSspec under the same adaptation setting, and ZS and FT for the same input source. Asterisks indicate statistically significant differences after correction for multiple comparisons ($P_{\mathrm{adj}} < 0.05$).

A Spectral validation of IFSspec

The paired 850-hPa kinetic energy spectra show that IFSspec retained the broad spectral shape of IFS while strongly reducing its short-wavelength power toward the AIFS spectrum (Fig. A1).

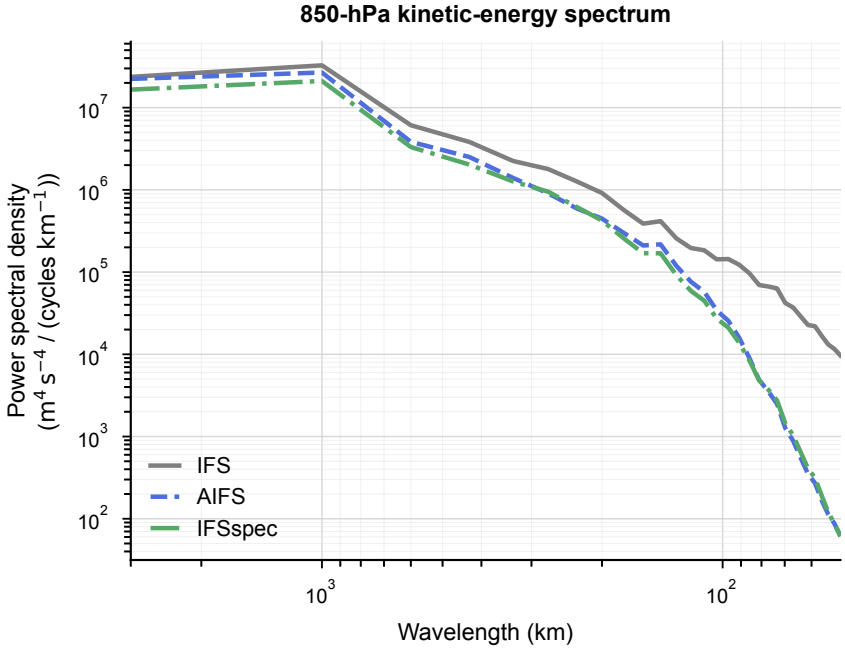


Figure A1. 850-hPa kinetic energy spectra. Curves show the mean one-dimensional isotropic kinetic energy spectra calculated from TC-centered fields common to the three datasets during 2023–2024 at forecast lead times of 24, 48, 72, 96, and 120 h.