Enhancing quantitative precipitation estimation of NWP model with fundamental meteorological variables and Transformer based deep learning model

Haolin Liu[1], Jimmy Chi-Hung Fung[1], Alexis Kai-Hon Lau[1], and Zhenning Li[2]

[1]Hong Kong University of Science and Technology
[2]Division of Environment and Sustainability, The Hong Kong University of Science and Technology

August 12, 2023

Abstract

Quantitative precipitation forecasting in numerical weather prediction (NWP) models rely on physical parameterization schemes. However, these schemes involve considerable uncertainties due to limited knowledge of the mechanisms involved in the precipitating process, ultimately leading to degraded precipitation forecasting skills. To address this issue, our study proposes using a Swin-Transformer based deep learning (DL) model to quantitatively map fundamental variables solved by NWP models to precipitation maps. Our results show that the DL model effectively extracts features over meteorological variables, leading to improved precipitation skill scores of 21.7%, 60.5%, and 45.5% for light rain, moderate rain, and heavy rain, respectively, on an hourly basis. We also evaluate two case studies under different driven synoptic conditions and show promising results in estimating heavy precipitation during strong convective precipitation events. Overall, the proposed DL model can provide a vital reference for capturing precipitation-triggering mechanisms and enhancing precipitation forecasting skills. Additionally, we discuss the sensitivities of the fundamental meteorological variables used in this study, training strategies, and performance limitations.

Enhancing quantitative precipitation estimation of NWP model with fundamental meteorological variables and Transformer based deep learning model

Haolin Liu[1], Jimmy C.H. Fung[1,2], Alexis K.H. Lau[1,3], Zhenning Li[1]

[1]Division of Environment and Sustainability, Hong Kong University of Science and Technology, Hong Kong, China
[2]Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong, China
[3]Department of Civil and Environmental Engineering, Hong Kong University of Science and Technology, Hong Kong, China

Key Points:

- We employ a transformer based deep learning model to improve the accuracy of precipitation estimation in NWP models.
- Various training strategies were implemented to manage the highly skewed precipitation data leading to improvements in heavy rainfall events.
- Evaluation conducted with multiple metrics including skill score, quantile and spatial distribution as well as two case studies.

Corresponding author: Jimmy C.H. Fung, `majfung@ust.hk`

Abstract

Quantitative precipitation forecasting in numerical weather prediction (NWP) models rely on physical parameterization schemes. However, these schemes involve considerable uncertainties due to limited knowledge of the mechanisms involved in the precipitating process, ultimately leading to degraded precipitation forecasting skills. To address this issue, our study proposes using a Swin-Transformer based deep learning (DL) model to quantitatively map fundamental variables solved by NWP models to precipitation maps. Our results show that the DL model effectively extracts features over meteorological variables, leading to improved precipitation skill scores of 21.7%, 60.5%, and 45.5% for light rain, moderate rain, and heavy rain, respectively, on an hourly basis. We also evaluate two case studies under different driven synoptic conditions and show promising results in estimating heavy precipitation during strong convective precipitation events. Overall, the proposed DL model can provide a vital reference for capturing precipitation-triggering mechanisms and enhancing precipitation forecasting skills. Additionally, we discuss the sensitivities of the fundamental meteorological variables used in this study, training strategies, and performance limitations.

Plain Language Summary

Numerical weather prediction (NWP) models depend on certain empirical formulations known as parameterizations to estimate precipitation. However, these methods often fall short due to the intricate dynamics of rainfall, which involves numerous small-scale interactions that these models are unable to fully capture. To counteract these limitations, our study deploys a form of machine learning known as deep learning (DL) to predict precipitation. This DL model utilizes fundamental weather variables derived from NWP models to make its estimations, serving as a remedy for the inherent weaknesses of traditional models caused by the uncertainties in their parameterization schemes. The implementation of our deep learning model resulted in a significant enhancement in rainfall prediction accuracy, particularly in the case of extreme precipitation events. This suggests that the application of machine learning strategies could be a promising approach to improve the reliability of rainfall forecasts, a crucial element for effective weather prediction and water resource management.

1 Introduction

Accurate quantitatively forecasting precipitation is essential for future planning and very helpful for minimizing human lives and property damage beforehand of extreme events, especially under the current rapidly changing climate. NWP models have been playing an increasingly important role in all operational centres and academia for understanding our earth system. It relies on discretizing a full set of governing equations including the Navier-Stokes equations, ideal gas law and thermodynamics and solving them numerically (Kalnay, 2003). With computing these prognostic equation sets over physical grids across different scales, the spatiotemporal evolving of the meteorological variables such as the temperature, wind speed and direction, air pressure and density are represented under the rotating earth coordinates. Building upon the enrichment of scientific knowledge in fundamental physics, and accelerated with the advances in technology such as computational power and numerous sources of observational data, NWP is showing quite a revolution over the past decades (Bauer et al., 2015).

As a result, the forecasting of fundamental meteorological variables under a so-called resolved scale of motion is readily available and more reliable in terms of its accuracy. However, many processes under the unresolved scales of motion also enter the equations, such as the moist processes involving condensation and evaporation, turbulence, convective activities and cloud microphysics, which are tightly related to precipitation formation and need to be parameterized to describe their relations with the states in resolved scales (Bauer et al., 2015). These parameterization schemes are generally based on the simplification and approximation of the physic laws to facilitate the numerical solutions, hence carefully chosen and sensitivity tests for the parameterization schemes will considerably affect the precipitation forecasting skills. Moreover, with an insufficient understanding of the underlying physics and some inherent uncertainties, using parameterization schemes will intrinsically bottleneck the further improvements of the performance for quantitatively estimating the precipitation (Zhou et al., 2022).

With the blossoming of artificial intelligence, many researchers have demonstrated the great ability of deep learning models in handling geoscience and remote sensing tasks, including precipitation estimation and forecasting. Shi et al. (2015) proposed and evaluated a series of spatial-temporal models dealing with precipitation nowcasting problems by extrapolating radar echoes and achieved better performance compared to traditional optical flow method (Shi et al., 2015, 2017). Ravuri et al. (2021) proposed to use a generative model with the stochastic method to extend nowcasting leading time without resorting to blurring. Sønderby et al. (2020) constructed a deep learning predictive model that uses satellite, radar and precipitation data and achieved a forecast leading time of 8 hours with a high spatiotemporal resolution, and outperforms the High-Resolution Rapid Refresh (HRRR) in terms of its accuracy. Other than precipitation nowcasting tasks, machine learning tools are also commonly applied to satellite images for precipitation estimation. Tao et al. (2017) proposed a deep learning model to extract features from bispectral satellite infrared (IR) and water vapour (WV) channels for detecting rain areas. Chen et al. (2019) proposed a two-stage hybrid neural network to estimate precipitation using ground-based radar and satellite observations. Wang et al. (2021) proposed a transfer learning based method, which uses data-riched Continental US (CONUS) IR dataset from the Geostationary Operational Environment Satellite (GOES) for pre-training of the model, and then transferred to China through re-training with multi-band IR signals from Chinese Fengyun (FY) satellite. Gao et al. (2022) used a U-Net model combined with the attention mechanism to directly retrieve precipitation maps using multi-band FY satellite images at a near real-time scale.

Many recent studies attempt to use data-driven models directly perform the NWP tasks in favour of their computational efficiencies compared to state-of-the-art NWP models. These data-driven models are generally trained on climate model outputs, general circulation models (GCM)(Scher & Messori, 2019; Chattopadhyay et al., 2020), or trained on reanalysis products such as ECMWF Reanalysis v5 (ERA5) dataset (Rasp et al., 2020; Rasp & Thuerey, 2021). Dueben and Bauer (2018) presented a “toy model” to identify challenges and fundamental design choices for deep learning based forecasting systems. Arcomano et al. (2020) designed a deep learning model and performed a 20-day global forecast. Evaluation results indicate that the DL model outperforms the NWP models for those state variables most affected by parameterization processes. Other than convolutional based deep learning models, Pathak et al. (2022) built a Fourier operator based transformer network to perform weather forecasting at globally 0.25° resolution and achieved matched accuracy with the state-of-the-art NWP model system the ECMWF Integrated Forecasting System (IFS).

Another prominent application of machine learning and deep learning techniques for NWP tasks is post-processing and bias correction. Grönquist et al. (2021) applied a convolutional neural network for bias correction of ensemble NWP predicted temperature field at various pressure levels, and achieved 14% improvement of ensemble forecast skill (CRPS) with a considerable reduction of computational cost owing to reduce the usage of trajectories. Taillardat and Mestre (2020); Li et al. (2022); Hess and Boers (2022) used machine learning and deep learning frameworks for post-processing quantitative precipitation forecasting results on the ensemble NWP models and achieved promising results on estimating heavy rainfall events located at long tails of the distribution curve.

With many machine learning methods have achieved remarkable results for nowcasting tasks or forecasting basic meteorological variables mentioned above. Directly map-

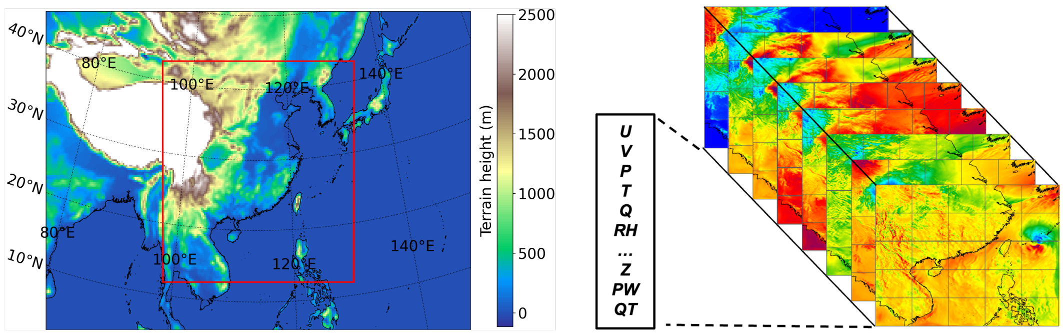

Figure 1. Left: Study domain and terrain height for input WRF meteorological data. Right:Input meteorological data simulated by WRF model.

ping basic meteorological variables to precipitation amounts using a machine learning model instead of parameterization schemes has been rarely explored. Therefore, in this study, we aim to develop a deep learning method for extracting rainfall features, from basic meteorological variables including temperature, water vapour, and atmospheric movements simulated by Weather Research and Forecasting Model (WRF) model at 27-km resolution. The basic variables were fed to an attention mechanism based Shift Window Vision Transformer (Swin Transformer) neural network (Dosovitskiy et al., 2021; Liu et al., 2021), and targeted to reproduce the high-resolution satellite rainfall product, the Climate Prediction Center morphing method (CMORPH) data (Xie, Pingping et al., 2019). This deep learning method will circumvent uncertainties of the physical parameterization scheme owing to incompletely understood physical processes and capture the non-linearity relationship between the predictors and labels.

2 Methodology

We consider the task of quantitative precipitation estimation using a deep learning model as an optimization problem, which can be formulated as follows:

$$\boldsymbol{\Theta} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \int \mathcal{L}(\boldsymbol{\Psi}(\boldsymbol{X};\boldsymbol{\theta}), \boldsymbol{y}) dy \tag{1}$$

In this formulation, the model takes pairs of input consisting of basic atmospheric variables $\boldsymbol{X}$ and precipitation observational data $\boldsymbol{y}$. The mapping function $\boldsymbol{\Psi}$ is used to relate these inputs, and it has trainable parameters $\boldsymbol{\theta}$. The goal is to find the optimal parameters $\boldsymbol{\Theta}$ by minimizing a set of loss functions $\mathcal{L}$ using optimization algorithms.

2.1 Dataset preparing

2.1.1 Trainning predictors:

To generate a dataset for high-resolution precipitation maps, we conducted a long-term dynamical simulation of 5 years ($2017 - 2021$) over the wettest season of south-east China, as shown in the left panel in Fig.1. The simulation was performed using the Weather Research and Forecasting (WRF) model with driven data from ERA5. The raw resolution for ERA5 data is $0.25^{\circ} \times 0.25^{\circ}$, and we used the meteorological fields from the WRF domain 1 simulation with 27-km resolution as predictors.

For the numerical simulation, we used a vertical grid with 38 levels to accurately represent the atmospheric system. We selected four model layers of three-dimensional

(3D) basic variables, including wind velocity (U, V, W), pressure (P), temperature (T), geopotential height (z), and humidity (Q), and stacked them with the corresponding variable at the surface level. We also included a two-dimensional (2D) diagnostic variable, total precipitable water (TPW), in our analysis. The vertical wind speed (W) at a height of 10 meters was not available in our model output, so the number of layers for this variable is limited to four. These variables were combined to form a 34-layer feature map, which we will refer to as 34 channels, as shown in the right panel Fig.1.

In atmospheric modelling, it is common practice to use pressure levels due to the decrease in pressure with height in the atmosphere. However, interpolating values from the model layers to the pressure levels can sometimes result in missing values of the basic variables due to varying terrain heights. To avoid this issue and minimize the memory requirements for training a deep learning model, we chose to directly use the values from the model layers rather than interpolating to pressure levels. This allowed us to accurately represent the atmospheric system while minimizing the computational resources needed for the simulation.

2.1.2 Observational precipitation data:

To obtain observational precipitation data as the reference for ground truth, we used CMORPH (NOAA CPC Morphing Technique), a high-resolution global satellite precipitation product. The data is created by combining passive microwave and infrared wave radiance measurements from multiple satellite instruments and adjusted using daily rain gauge analysis. The full-resolution CMORPH data used in this study has a high spatial resolution of 8 km and a frequency of 30 minutes. To align it with the meteorological data from the model used in this study, the CMORPH data was resampled to an hourly frequency, and its pixel values were matched to the corresponding grid points in the model.

2.2 Network Architecture

The architecture of the proposed model shown in Fig.2 A is based on the classical encoder-decoder framework, which has been successfully applied to many semantic segmentation tasks in the computer vision field. The model is inspired by the original UNet model (Ronneberger et al., 2015). The gridded meteorological data generated by the WRF model are first divided into 4×4 non-overlapping patches by using a 2D convolutional layer with a stride and kernel size equal to the patch size. The patches are then transformed into sequence embeddings and fed into the encoder.

In the encoder, we replace the CNN backbone network in the original UNet model with the Swin-transformer (Liu et al., 2021). Each encoder block consists of a patch merging layer and four Swin-transformer blocks. The patch merging layer performs downsampling, similar to the pooling operation in CNN-based models, while the Swin-transformer block extracts features, similar to a convolutional operation. The input data passed through the patch partition layer has a size of $\frac{H}{4} \times \frac{W}{4}$ with C embedding channels, where H and W are the height and width of the input data. With each pass through an encoder block, the height and width are halved and the number of channels doubles.

In the decoder block, the patch-expanding layer performs upsampling using bilinear interpolation to restore the feature map to its original resolution, doubling the size of the feature map and reducing the feature dimension to half of its original dimension. To maintain the information lost during downsampling in the encoder, the expanded feature maps are fused with the downsampled features through a skip connection structure. This allows the Swin-transformer blocks in the decoder to receive inputs with the same size as the corresponding level of the encoder, but with features crossing multiple dimensions.

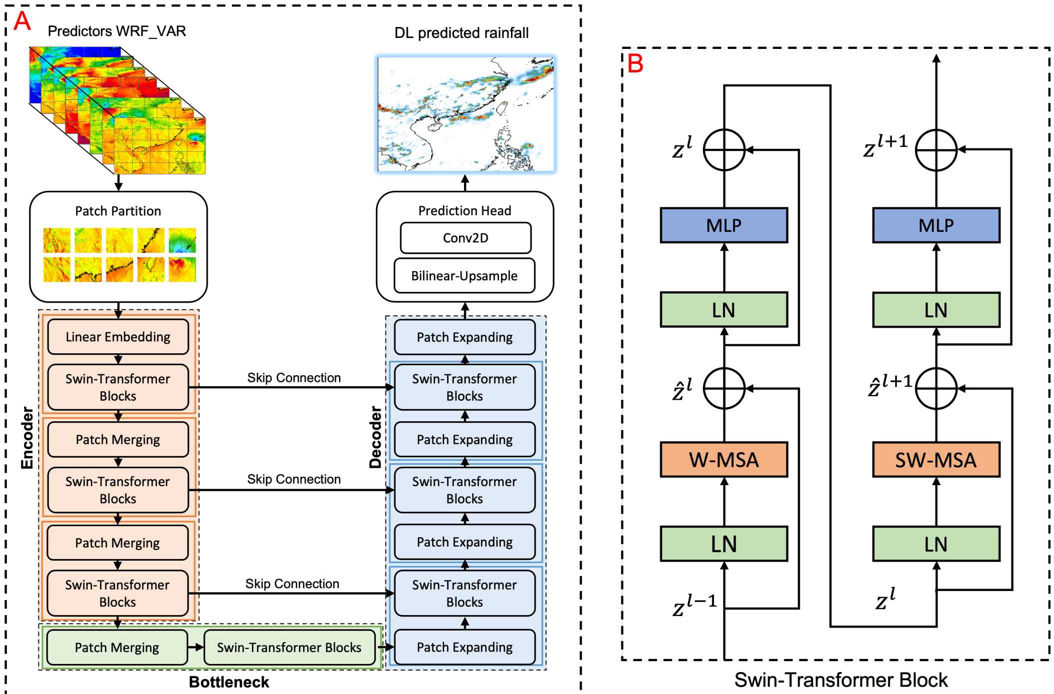

Figure 2. Architechture of the Swin-Transformer-Unet model (A), and the basic computational principle for a Swin Transformer block (B).

The bottom levels of the encoder and decoder are connected by a bottleneck, which has the same structure as the encoder but only with two Swin-transformer blocks. This hierarchical architecture design is rooted in the principle of enhancing the network's ability to learn features at multiple scales, which is particularly important for meteorological data due to its multi-scale nature.

The computation route of the Swin-transformer block, as depicted in Fig.2 B, is designed to reduce the computational cost compared to traditional multi-head self-attention (MSA) modules through the implementation of a sliding window mechanism. The Swin-transformer structure (Liu et al., 2021), accomplishes this reduction by composing a Swin-transformer block of two consecutive attention modules. Each attention module is composed of two LayerNorm (LN) layers, a multi-head self-attention module, a residual connection shortcut, and a 2-layer multilayer perceptron (MLP) with Gaussian Error Linear Units (GELU) nonlinearity.

Before the MSA module and the MLP module, the LayerNorm (LN) is applied, and the two attention modules differ in the type of multi-head self-attention employed. The first MSA layer uses a regular Window-Based MSA (W-MSA), while the second MSA layer adopts a Shifted Window-Based MSA (SW-MSA) module. The calculation of the Swin-transformer block is described by the following equations:

$$\hat{z}^l = W\text{-}MSA\left(LN\left(z^{l-1}\right)\right) + z^{l-1} \tag{2}$$

$$z^l = MLP\left(LN\left(\hat{z}^l\right)\right) + \hat{z}^l \tag{3}$$

$$\hat{z}^{l+1} = SW\text{-}MSA\left(LN\left(z^l\right)\right) + z^l \tag{4}$$

$$z^{l+1} = W\text{-}MSA\left(LN\left(\hat{z}^{l+1}\right)\right) + \hat{z}^{l+1} \tag{5}$$

where $\hat{z}^l$ is the output features of the (S)W-MSA module, z^l is the output features of the MLP module, and l represents the number of blocks.

2.3 Experiment details

2.3.1 Loss functions:

In meteorology, most variables, such as temperature and wind speed, are typically represented as continuous values over the model grids. These variables are usually smooth and evenly distributed unless there are significant changes in terrain or extreme weather conditions. Precipitation data, however, is often underestimated during extreme precipitation events due to its imbalanced distribution, making it a challenging problem to accurately predict from an engineering perspective. To address this issue, we propose the use of the Tversky (Salehi et al., 2017) loss function in our model. The Tversky loss function is defined as:

$$\mathcal{L}_{Tversky} = 1 - \frac{\sum_i^N p_i g_i}{\sum_i^N p_i g_i + \alpha \sum_i^N p_i(1-g_i) + \beta \sum_i^N (1-p_i) g_i} \tag{6}$$

where p_i and g_i are the predicted and ground truth values at pixel i, respectively, and α and β are the weighting factors for the false positive and false negative terms. It can be viewed as a generalization version of the Dice similarity coefficient, which is widely used in image segmentation tasks. The advantage of using the Tversky loss function is that it provides better control over the trade-off between precision and recall by allowing for the adjustment of the false positive and false negative weighting factors. This is particularly useful in our situations where the precipitation data is extremely imbalanced, with the background pixels represented by 0 indicating no rainfall. To encourage the model to better predict extreme precipitation, the Tversky loss function has been adjusted to assign higher penalties to false negative predictions. This is achieved by setting β to a higher value than α in our experiments, making the model more inclined to accurately predict extreme precipitation values. In addition to the Tversky loss function, we also use the mean squared error (MSE) loss function, which is a commonly used loss function for regression problems, as our precipitation prediction task can also be viewed as a pixel-wise regression problem. The combined loss function can be expressed as:

$$\mathcal{L}(y, \hat{y}) = 0.5\mathcal{L}_{Tversky}(y, \hat{y}) + 0.5\mathcal{L}_{MSE}(y, \hat{y}) \tag{7}$$

2.3.2 Data transform:

To enhance the network's convergence, we introduce data transformations to the original input data. Initially, a center-cropping operation is executed on the meteorological data to mitigate boundary errors. This step is crucial as the WRF model, being a regional weather forecasting model, necessitates input from the global background field from ERA5. This requirement can potentially introduce errors at the lateral boundaries of the study domain. Next, we apply mean-std normalization to the input data to scale each predictor. This helps to bring the data into a similar range and prevent one predictor from dominating the other. Finally, we take a log transformation of the observational label data, the precipitation map, to reduce the skewness of its distribution. With the log transformation, the distribution of data is better represented and improves the convergence in the training as well as the accuracy while inferencing.

2.3.3 Evaluation metrics:

To evaluate the performance of the model, various metrics are calculated including the probability of detection (POD), threat score (TS), equitable threat score (ETS), false alarm ratio (FAR), and BIAS. These metrics are defined as follows:

$$POD = \frac{h}{h + miss} \tag{8}$$

$$TS = \frac{h}{h + miss + f} \tag{9}$$

$$ETS = \frac{h - h_{random}}{h + miss + f - h_{random}} \tag{10}$$

$$BIAS = \frac{h + f}{h + miss} \tag{11}$$

$$FAR = \frac{f}{h + f} \tag{12}$$

The POD primarily focuses on the number of hits, while the TS, FAR, and BIAS evaluate the combined impact of hits and false alarms. The ETS takes into account the possibility of a hit by chance by calculating $h_{random} = \frac{h+f}{h+m+f+cn}$, where cn is the number of correct negatives. A higher POD, TS, and ETS, and a lower FAR or a BIAS closer to 1 are considered an indicator of a more accurate prediction.

3 Analysis of experiment results

3.1 Overall performance

In our study, we utilized a neural network that was trained using 6 years of WRF simulation data, and the WRF simulation was driven by the ERA5 reanalysis data, which is the foundation of our deep-learning model. Therefore, the performance of our deep-learning model is ultimately dependent on the original ERA5 background and our WRF simulation. Despite the advancements in weather forecasting, it remains challenging to reproduce the heavy precipitation events that are generated by intense convective systems. By utilizing the Swin-transformer Unet to process fundamental meteorological variables, we observed significant improvements in the prediction of heavy precipitation events on the tail, both in terms of accuracy in rainfall quantities and the location of the rainfall. The overall evaluation scores calculated from July to December 2021, with all leading times from 0 to 72 hours, are listed in Table 1.

(1) Our deep learning model for the prediction of drizzles with a rainfall intensity greater than $1mmh^{-1}$ has been shown to enhance POD, TS, and ETS, while simultaneously reducing the FAR. Additionally, the FBIAS is closer to 1 when compared with the results obtained from a pure WRF simulation. The improvement in the TS score is particularly significant, reaching as high as 21.7%.

(2) In the prediction of moderate rainfall with an hourly intensity of 3mmh^{-1} and 5mmh^{-1}, the POD has increased from 0.145 to 0.218 and from 0.088 to 0.161 respectively, while the FAR has slightly decreased. As a result, the TS and ETS scores have also seen considerable increases, with the TS score increasing from 0.117 to 0.164 for the 3mmh^{-1} threshold and from 0.076 to 0.122 for the 5mmh^{-1} threshold. The ETS score has increased from 0.106 to 0.151 for the 3mmh^{-1} threshold and from 0.071 to 0.114 for the 5mmh^{-1} threshold. The relative improvement ratio for both the TS and ETS scores is as high as 60.5% for the 5mmh^{-1} threshold.

(3) For the heavy rainfall events with an hourly precipitation intensity exceeding 10mm, our model has demonstrated the ability to significantly enhance the POD, TS, and ETS scores. However, in detecting heavy rainfall, there is a trade-off of introducing higher FAR. This indicates that the meteorological background data may not fully match the observational precipitation data, causing some intensive weather systems simulated by the WRF model with possibilities of heavy precipitation to be mistakenly placed. The improvement in POD for heavy rainfall, with the addition of both 10mm and 20mm intensities, is around 50%, increasing from 0.06 to 0.12. Similarly, the improvement in the TS and ETS scores ranges from 0.055 to 0.08, with an improvement rate of 45.5%. The pure WRF simulation has a BIAS of around 0.1 for heavy rainfall, which suggests that the detected rainfall area is sub-optimally small. Our deep-learning model improved

Table 1. The evaluation results for the WRF simulation and WRF + Swin-transformer Unet for the hourly rainfall intensity with forecasting leading time of $0 - 72h$, the evaluation period is July to December 2021. Evaluation metrics including POD (probability of detection), FAR (false alarm ratio), TS (threat score), ETS (equitable threat score), and BIAS ratio are listed.

	Precipitation (mmh^{-1})	**POD**	**FAR**	**TS**	**ETS**	**FBIAS**
WRF	0.1	0.508	0.568	0.295	0.213	1.263
	1.0	0.326	0.611	0.209	0.172	0.902
	3.0	0.145	0.610	0.117	0.106	0.378
	5.0	0.088	0.605	0.076	0.071	0.223
	10.0	0.045	0.563	0.040	0.039	0.123
	20.0	0.016	0.337	0.014	0.014	0.073
WRF + AI	0.1	0.525	0.475	0.359	0.289	1.013
	1.0	0.346	0.523	0.254	0.222	0.726
	3.0	0.218	0.611	0.164	0.151	0.551
	5.0	0.161	0.667	0.122	0.114	0.462
	10.0	0.086	0.755	0.063	0.061	0.391
	20.0	0.030	0.695	0.021	0.020	0.356

this score to nearly 0.4, demonstrating a better ability to detect heavy rainfall. Improving the accuracy of heavy rainfall detection and reducing the FAR is an important area for further research.

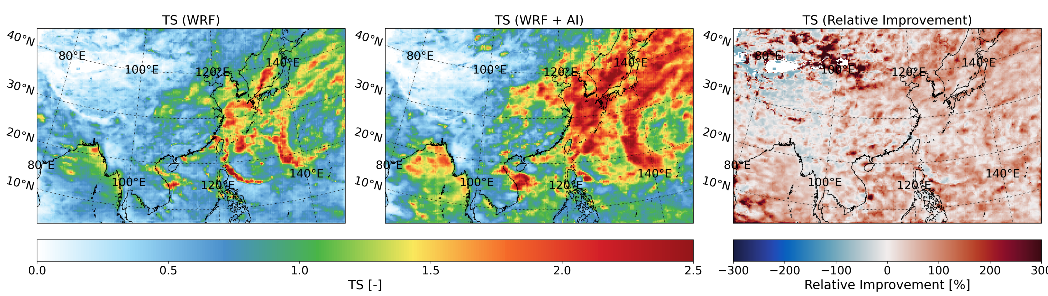

Figure 3. Spatial distribution of Threat Score (TS) for the baseline WRF, WRF + AI framework, and its relative improvements over the evaluation dataset.

(4) In terms of spatial distribution, our deep learning framework exhibits a substantial enhancement in forecast skill, as quantified by the cumulative TS score across all precipitation intensity thresholds. Notably, this improvement is observed in areas prone to heavy precipitation events, such as potential tropical cyclone pathways and associated rainbands, in addition to Mei-Yu frontal systems during monsoon seasons. The relative advancements in these regions amount to several-fold increases, while minor degradation is detected in an insignificant fraction of pixels, ensuring overall model performance remains robust mesoscale application.

3.2 Quantile distribution

Quantile distribution stands as a pivotal measure for evaluating the predictive proficiency of heavy rainfall events. Fig. 4 demonstrates a monotonic decline in model per-

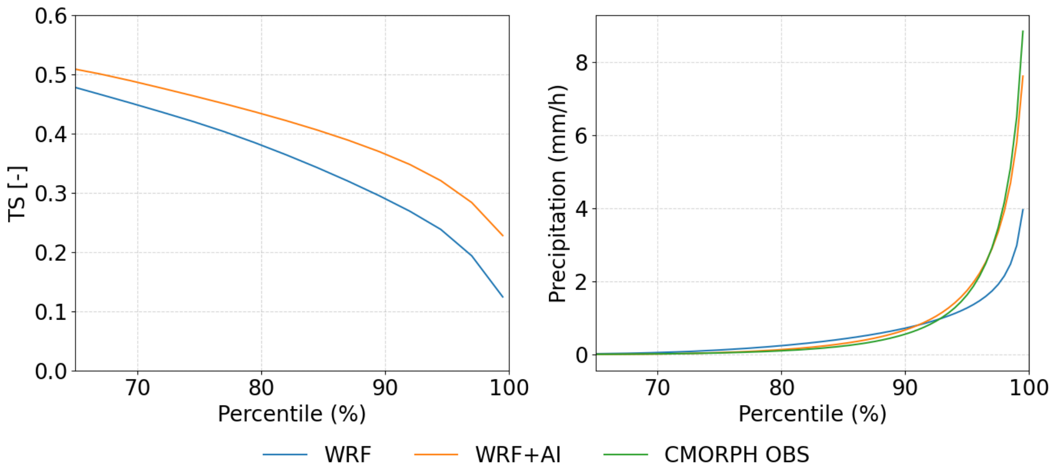

Figure 4. Threat Score (TS) for precipitation events above the percentile thresholds for WRF and WRF + AI framework (left), and hourly precipitation intensity for WRF, WRF + AI, CMORPH observation at each corresponding percentiles.

formance across varying quantile thresholds, signifying the escalating challenge of accurately predicting more extreme events.

When examining lower percentiles, our baseline WRF simulation, which employs the Kain-Fritsch (KF) cumulus parameterization scheme (Kain & Fritsch, 1993; Kain, 2004), displays performance characteristics akin to our deep learning framework in terms of its TS, albeit with a slight overestimation of light rainfall between the $70th$ and $90th$ percentile with an hourly rainfall intensity less than $1mm$. This is primarily due to the fact that the KF scheme is well-suited to accommodate light precipitation events, which are less intricate to predict. This finding is consistent with several dynamic downscaling studies (Ma & Tan, 2009; Otieno et al., 2020) that have indicated a tendency of the KF scheme to produce a wet bias during light rainfall while demonstrating limited predictability for heavier rainfall events.

When we examine more intense rainfall events that surpass the $95th$ percentile threshold, our proposed deep learning framework begins to exhibit superiority in its capacity to replicate these extreme events and align the distribution more closely with observed ground truth. Within this percentile range, the TS score for the deep learning framework is approximately double that of the baseline WRF model, and the estimated rainfall intensity is closely tracks the CMORPH observation until the 99.5th extreme percentile is reached. At this point, the largest bias is about 20%, whereas the baseline WRF model can only capture less than 50% of the intensity. This result underscores the enhanced performance of this deep learning framework in the accurate prediction and representation of intense rainfall events.

In addition to the above, we present the meridional and zonal averages of hourly precipitation intensity at the $95th$ percentile in Fig. 5, as well as its spatial distribution in Fig. 6. The application of our deep learning framework enables us to refine the estimated rainfall intensity at the $95th$ percentile, resulting in a significantly improved alignment with the CMORPH observation, both spatially and in terms of its zonal and meridional mean. The spread property, represented by the shadowed area, also aligns more closely with the observations.

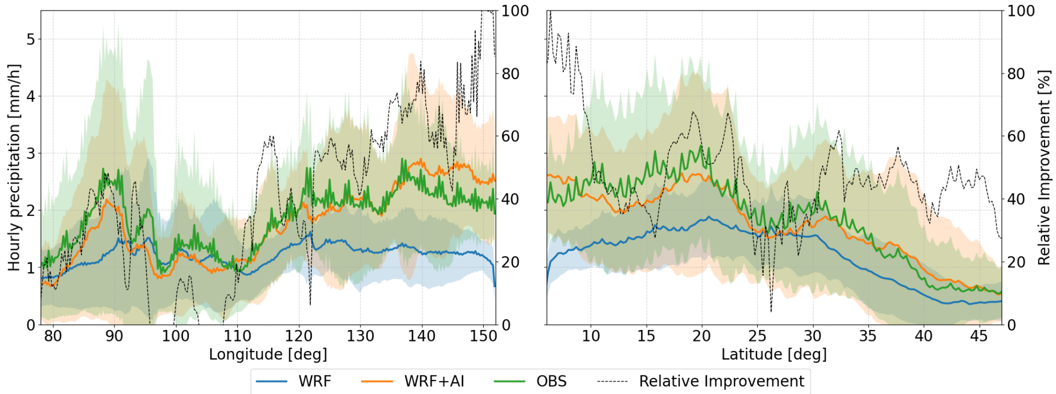

Figure 5. Meridionally (left) and Zonally (right) averaged precipitation intensity at 95th percentile.

The highest relative improvements are evident in the eastern part and southwest quadrant of our experimental domain, with the maximum relative improvement exceeding 100%. Some minor fluctuations and degradation are noticeable around 100°−110° Longitude and approximately 25° Latitude, the mountainous southwest part of China. This suggests that our deep learning framework may demonstrate lower confidence when dealing with orographic precipitation. Similar observations can be made from the spa-

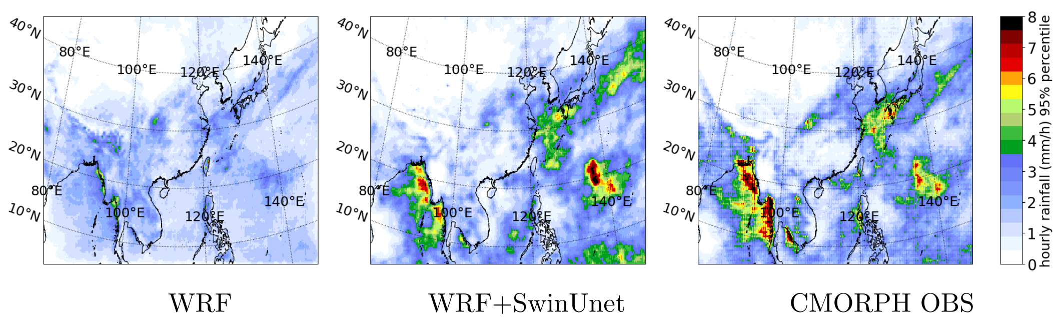

Figure 6. Spatial distribution of 95th hourly precipitation intensity for the baseline WRF (left), WRF + AI (mid) and CMORPH observation (right).

tial distribution plot in Fig. 6. The largest improvements are seen over potential tropical cyclone initiation areas in the western Pacific Ocean, the eastern China region influenced by the Mei-Yu frontal rain belts around the middle and lower reaches of the Yangtze River, the southern part of Japan and the Korean Peninsula, as well as the Southeast Asia region around the Bay of Bengal. Other regions also show varying degrees of improvement in terms of spatial distribution, indicating that our deep learning framework more accurately represents extreme precipitation patterns compared to the baseline WRF simulation, which exhibits limited predictability over these heavily precipitating areas during the wet season.

3.3 Case study

In this section, we present two case studies of heavy precipitation recorded in the study domain, as shown in Fig. 8 and Fig. 10. The precipitation maps depict the ac-

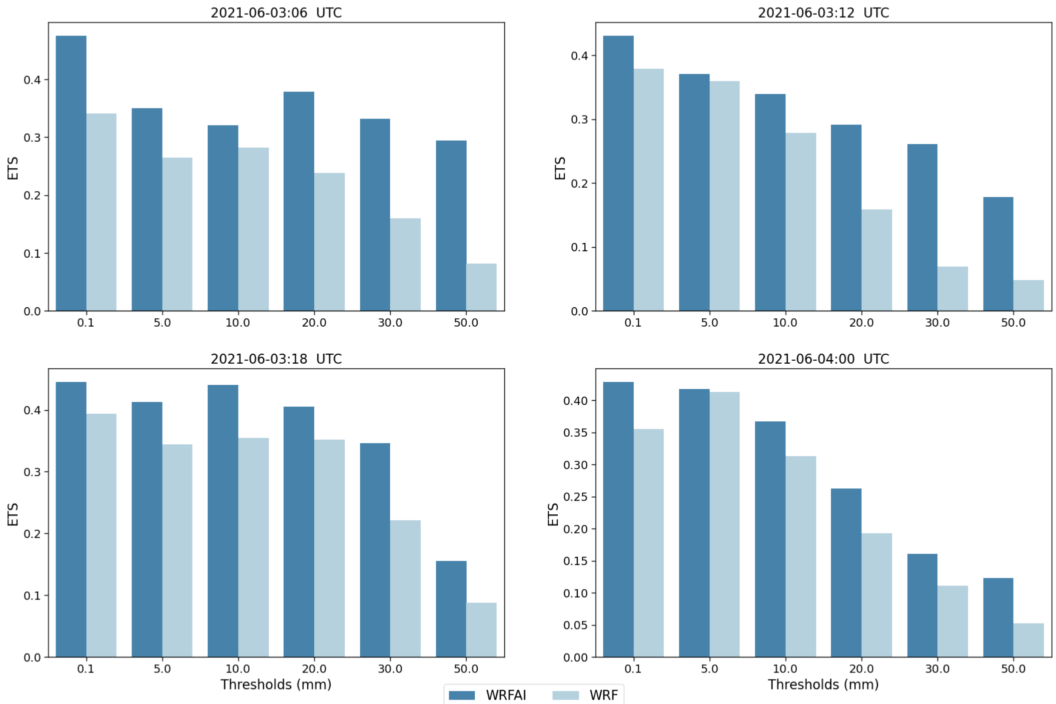

Figure 7. Equivalent Threat Score (ETS) of 6-hour accumulated precipitation for the case study on June 3rd, 2021 at 06:00 UTC, 12:00 UTC, 18:00 UTC, and 24:00 UTC

cumulated intervals of 6 hours and cover a period of one day for each case. The first case study occurred on June 3rd, 2021, while the second case study occurred on August 20th, 2021.

3.3.1 2021-06-03:

On June 3rd, 2021, a frontal rain band was observed to be moving southeastward. In our default WRF simulation, results are generally indicating weak signals for several precipitation hot spots over both the continent and the ocean side.

To improve the accuracy of our predictions, we utilized the Swin-Transformer-Unet model in conjunction with the basic meteorological variables predicted by WRF. As a result, the overall precipitation patterns observed are exhibiting more consistency with the CMORPH observational dataset for all four intervals during this case.

The frontal rain band were initialized around the centre of the study domain (Latitude $27°N$ Longitude $110°E$) before 06 UTC (Fig.8) and started moving southeastward driven by the low-pressure centre located on the northeast corner of the study area, the movement and the structure of the rain band is well preserved in the prediction results at 12 UTC by our deep learning model. In the following sequences of time, this frontal rain band further extended in length and almost covered the whole south and southeast part of China while approaching the coast.

Based on the quantitative evaluation shown in Fig.7, the performance of the default WRF model and the deep learning model were compared in terms of predicting drizzle and light rainfall with thresholds less than $10mm$ and heavier rainfall areas with thresholds exceeding $20mm$ or even $50mm$.

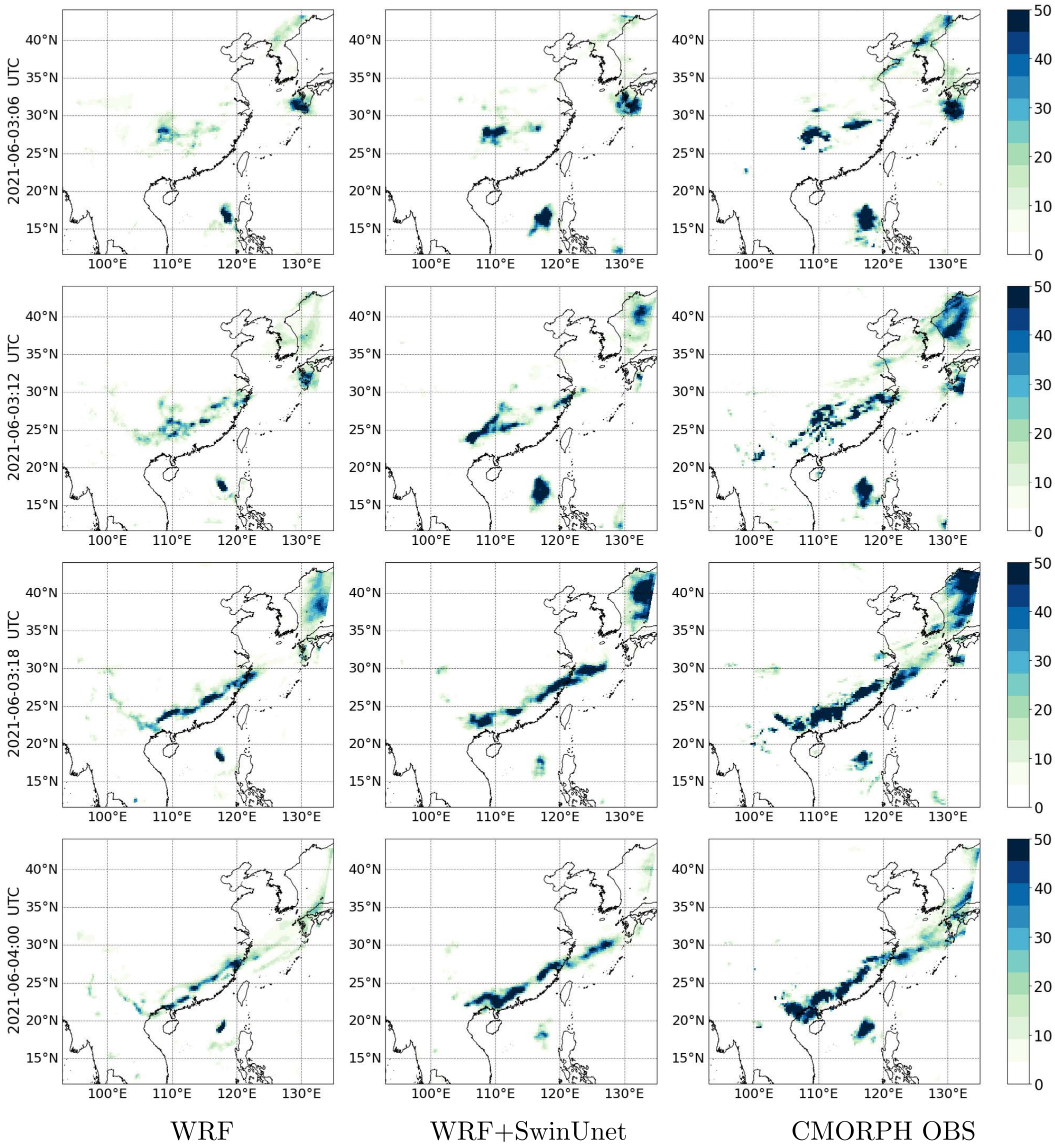

Figure 8. 6-h accumulated precipitation predicted by WRF, WRF + Swin-Transformer Unet, and observational data from CMORPH datasets on 2021-06-03.

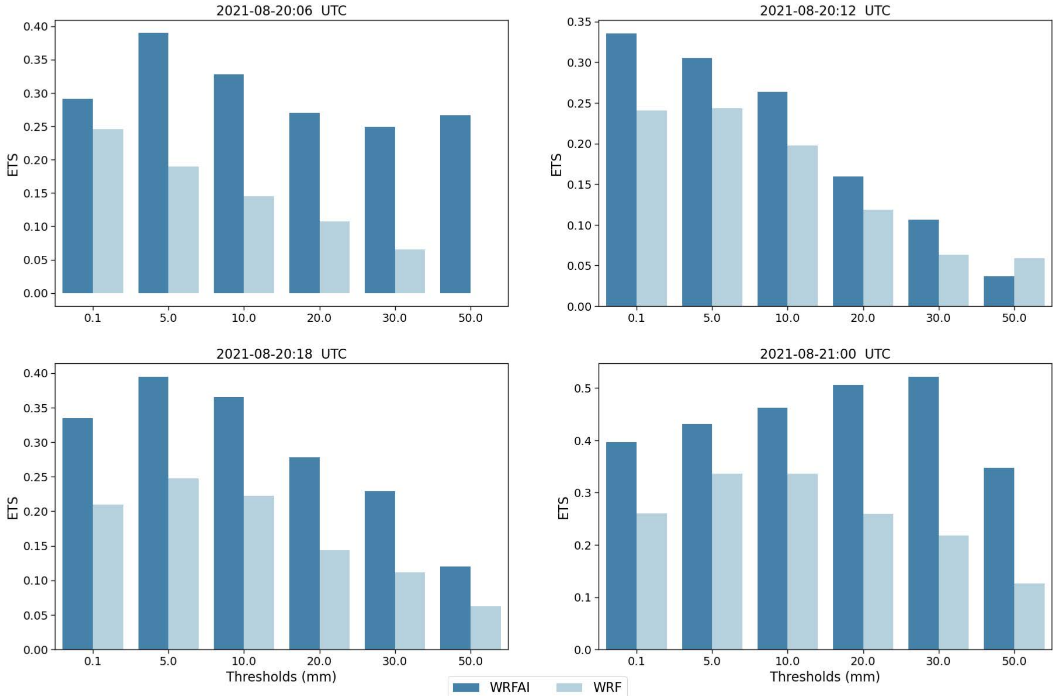

Figure 9. Equivalent Threat Score (ETS) of 6-hour accumulated precipitation for the case study on Aug 20th, 2021 at 06:00 UTC, 12:00 UTC, 18:00 UTC, and 24:00 UTC

For predicting drizzle and light rainfall with thresholds less than $10mm$ within a 6 hours interval, the ETS scores of both models were relatively close, except for the first time interval 06 UTC at $0.1mm$ thresholds, where the ETS score increased by nearly 40% from 0.34 to 0.5 in the deep learning model.

On the other hand, for heavier rainfall areas with thresholds exceeding $20mm$ or even $50mm$, the deep learning model outperformed the default WRF model by doubling or even trebling the ETS score, as indicated by the results at 06 UTC and 12 UTC. Moreover, the decrease of ETS for the deep learning enhanced prediction was less steep, which indicates a more stable performance in estimating precipitation for all ranges compared to the default WRF model.

It is worth noting that NWP models like the WRF model generally suffer from sharp degradation in performance when moving from the synoptic scale to the convective scale. Therefore, the superior performance of the deep learning model in estimating precipitation for heavier rainfall areas suggests its potential for improving the prediction accuracy of convective precipitation in NWP models.

3.3.2 2021-08-20:

On August 20th, 2021, Central and Northeast China experienced extreme precipitation, accompanied by thunderstorms and strong convective weather. Heavy rainfall of over 100mm was initially observed before 12:00 UTC (Fig.10) in the Central China region (located at Latitude $33°N$ and Longitude $115°E$). Subsequently, later in the day, extreme was observed in the Yellow Sea, Northeast China, and the Korean Peninsula.

As shown in the precipitation map in Fig.10, our baseline WRF simulation captures only a limited signal of the strong convective rainfall due to the coarse domain grid

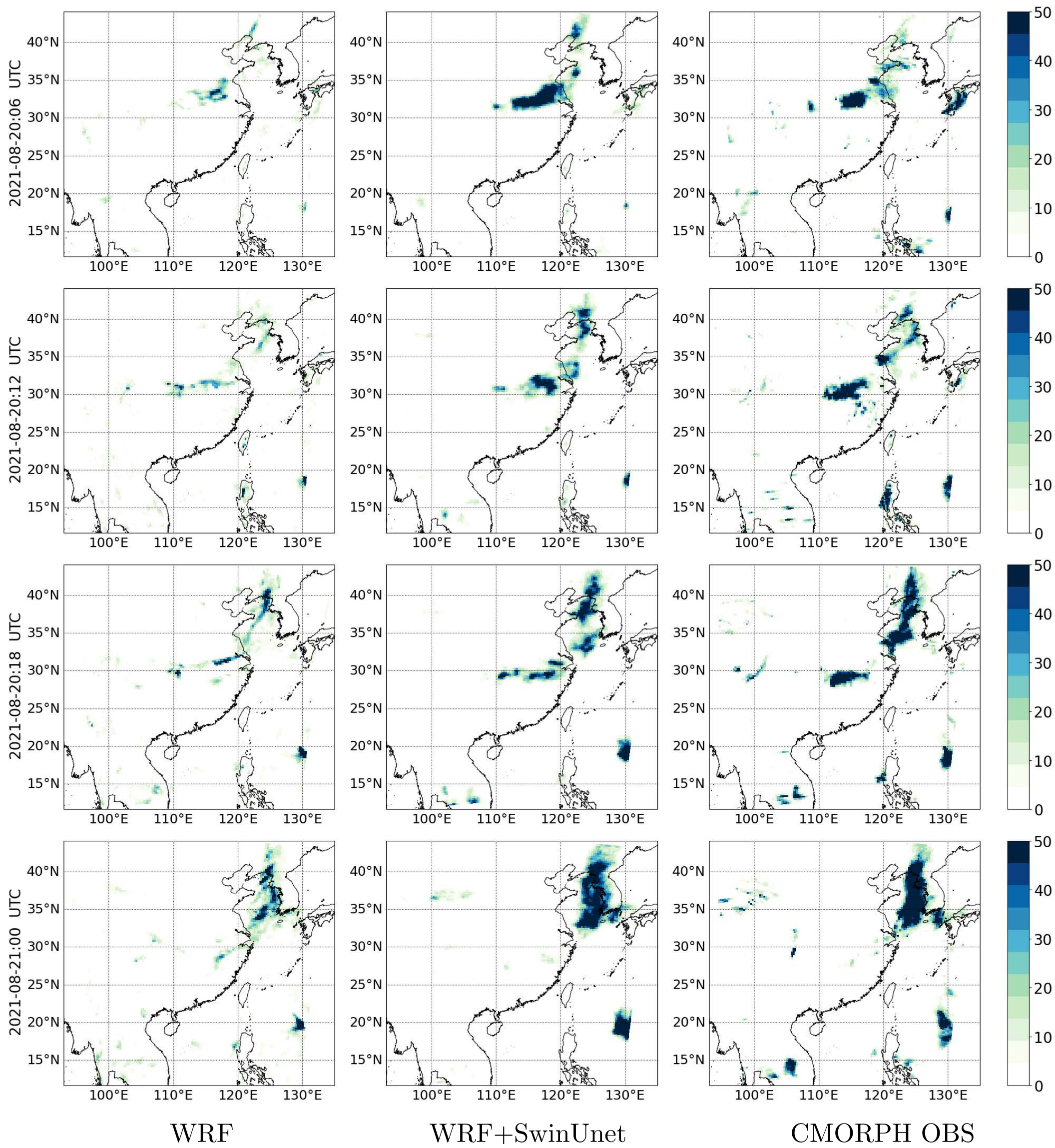

Figure 10. 6-h accumulated precipitation predicted by WRF, WRF + Swin-Transformer Unet, and observational data from CMORPH datasets on 2021-08-20.

size and the limitations of the parameterization scheme for cumulus and cloud microphysics. However, by enhancing the estimation of precipitation with our deep learning model, we were able to bridge the gap between the WRF simulation and the observations, caused by intrinsic limitations rooted in the parameterization schemes. This considerably reduces the negative bias and facilitates the estimation of extreme precipitation, both in its maximum precipitation amount and in reducing errors in its spatial distribution.

Fig.9 presents statistical evidence demonstrating the effectiveness of our deep learning framework. The results indicate that the most significant improvement over the baseline WRF simulation occurs at 20th 06 UTC and 21st 00 UTC, where the ETS score increases by an average of 30% for events with light to moderate rainfall (precipitation amount less than 10mm). Additionally, the baseline WRF simulation's performance degrades rapidly as the precipitation threshold increases, failing to detect precipitation exceeding 50mm in a 6-hour interval. In contrast, our deep learning model enhances the estimation results, maintaining a relatively good performance with an ETS score above 0.25 for all thresholds at 06 UTC and 00 UTC, while the baseline WRF model's performance drops to less than 0.1. These results indicate that our deep learning framework provides a significant improvement over the baseline WRF model, even though the grid size and parameterization schemes are not ideally suited for capturing strong convective precipitation during the monsoon season.

4 Conclusion and discussion

Regional numerical weather prediction (NWP) models like the WRF model are known to be sensitive to domain grid size (Jee & Kim, 2017) and parameterization schemes (Hasan & Islam, 2018), particularly when it comes to predicting precipitation. To address this challenge, a deep learning model for semantic segmentation using a Swin-Transformer backbone and a hierarchical Unet structure is proposed in this study. This model leverages basic meteorological variables such as air temperature, pressure, wind speed, and humidity to significantly improve the performance of the baseline WRF model in simulating precipitation, particularly for extreme events induced by strong convection. The overall effectiveness of this deep learning post-processing framework is demonstrated through a comprehensive performance evaluation, including an analysis of its spatial and quantile distributions, and a detailed discussion of two case studies.

To evaluate the model's performance, we assessed hourly precipitation amounts across intensity thresholds ranging from 0.1mm to 20mm during the period of June 2021 to September 2021. The results demonstrated that our deep learning model outperformed the baseline WRF simulation for all precipitation intensities. Specifically, the model improved the baseline WRF simulation by 21.7% for light rainfall and drizzle (precipitation amount less than $1mmh^{-1}$), and by 60% for moderate rainfall events with precipitation thresholds of $3mmh^{-1}$ and $5mmh^{-1}$. For heavy rainfall events with hourly precipitation intensity exceeding 10mm, the improvements reflected by the TS and ETS scores reached as high as 50% compared to the baseline WRF. The overall quantile distribution of baseline WRF and the proposed deep learning framework are also compared, with results showing that the prediction of rainfall intensity across all the quantiles received various degrees of improvement. Additionally, the spatial distribution of the $95th$ percentile rainfall intensity and its zonal and meridional averages were also revealing a significantly better alignment with observational data. However, minor challenges were noted in regions with possible orographic precipitation trigger mechanisms, particularly in the mountainous southwest part of China. Future exploratory efforts could be directed towards amplifying the model's proficiency in recognizing and integrating finer-scale terrain and land surface effects. Such advancements could potentially elevate the forecast skill in these currently less confident areas.

In addition to the overall evaluation, we presented two case studies of precipitation events triggered by different synoptic conditions to demonstrate the model's ability to capture complex weather phenomena. For both events, we investigated the 6-hourly accumulated rainfall of four intervals and showed how our deep learning model can provide more accurate precipitation forecasts by learning from meteorological datasets and extracting relevant features.

The first precipitation event is caused by the large-scale movement of frontal rain bands, while the second event is induced by strong convection during the monsoon season. In both cases, we observed rapid degradation of model performance as precipitation thresholds increased in the baseline WRF model. On the contrary, our deep learning model was able to compensate for the insufficient predictability of the baseline WRF model simulation and achieve improved ETS scores over each temporal interval at various precipitation thresholds. Notably, our model demonstrated particular success in capturing extreme precipitation amounts exceeding 30mm or 50mm, which are often difficult to predict using traditional modelling approaches. These findings demonstrate the effectiveness of our deep learning model in capturing precipitation characteristics from basic meteorological variables and further quantitively estimating precipitation based on extracted features.

As we are feeding the WRF model simulated meteorological fields into our deep learning model, the quality of the precipitation estimation results is ultimately dependent on the quality of our baseline WRF simulation, correspondingly, it should also be related to the initial forcing data used to drive the WRF simulation. Therefore, the results presented in this study is showing substantial relative improvements against the baseline WRF simulation, demonstrating the ability of this neural network in capturing triggering processes which are not currently described in the existing precipitation parameterizations of the WRF model. Moreover, due to the model abstraction, the model grid states may not fully match the CMORPH precipitation observations used as labels for training, making it challenging to accurately estimate precipitation amounts in terms of intensity and location. As a consequence, this spatial and temporal inconsistency of prediction and observation was reflected in the increased false alarm rate (FAR) at higher thresholds. Similar results were also noted by Hess and Boers (2022), they attributed this issue to the localized intermittent nature of the heavy rainfall events. We believe that this limitation can be mitigated by accumulating hourly prediction results over several hours, or by adjusting the hyperparameter α and β in the Tversky loss which controls the trade-off between precision and recall.

Compared to solely optimizing the model with a global MSEloss, several studies have attempted to manage the extremely skewed precipitation data by performing log-transformation scaling (Shi et al., 2017; Pathak et al., 2022), modifying traditional regression loss functions such as MAE loss and MSE loss by binning precipitation data and assigning different weights to each category (Shi et al., 2017; Franch et al., 2020), or combining structural similarity measure (SSIM) loss function during optimization (Tran & Song, 2019; Hess & Boers, 2022). The usage of the Tversky loss function has also demonstrated its superior ability in dealing with strongly imbalanced distributed precipitation data in our study. The current optimization process involves pixel-wise optimization of the Tversky loss function, followed by using MSE loss for global refinement. Exploring custom loss functions that can focus on local features and the overall distribution might be beneficial for further improving the temporal and spatial accuracy of the deep learning model.

Additionally, the data augmentation technique is also worth exploring. The existing permutation study by Li et al. (2022) has shown that the moisture-related predictor dominates the precipitation estimation in this deep learning framework. By using feature patch masking, mixing, and shuffling techniques, it may be possible to further improve the model's generalizing ability by increasing the difficulty of the original task and enriching the dataset's diversity. This approach can help reduce the model's reliance on specific predictors, leading to more robust and accurate predictions.

5 Open Research

The experiments are conducted using Pytorch software v2.0 (`https://pytorch.org/`). The precipitation data used for training as references are from CMORPH (`https://www.ncei.noaa.gov/products/climate-data-records/precipitation-cmorph`). The meteorological feature map used for training is generated from the WRF model (`https://github.com/wrf-model/WRF`), while due to the large size of the training dataset (∼300 G), it is currently archived in HKUST ENVF database (`http://envf.uswt.hk/itf-si/`), which can be provided upon request. The source code for preprocessing WRF data, training the Swin-Transformer-Unet model and post-processing to generate plots are available at (`https://doi.org/10.5281/zenodo.8210356`).

Acknowledgments

We appreciate the assistance of the Hong Kong Observatory (HKO), which provided the meteorological data. The work described in this paper was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project Nos. AoE/E-603/18 and T31-603/21-N).

References

Arcomano, T., Szunyogh, I., Pathak, J., Wikner, A., Hunt, B. R., & Ott, E. (2020, May). A Machine Learning-Based Global Atmospheric Forecast Model. *Geophysical Research Letters*, *47*(9). doi: 10.1029/2020GL087776

Bauer, P., Thorpe, A., & Brunet, G. (2015, September). The quiet revolution of numerical weather prediction. *Nature*, *525*(7567), 47–55. doi: 10.1038/nature14956

Chattopadhyay, A., Nabizadeh, E., & Hassanzadeh, P. (2020, February). Analog Forecasting of Extreme-Causing Weather Patterns Using Deep Learning. *Journal of Advances in Modeling Earth Systems*, *12*(2). doi: 10.1029/2019MS001958

Chen, H., Chandrasekar, V., Tan, H., & Cifelli, R. (2019, September). Rainfall Estimation From Ground Radar and TRMM Precipitation Radar Using Hybrid Deep Neural Networks. *Geophysical Research Letters*, *46*(17-18), 10669–10678. doi: 10.1029/2019GL084771

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N. (2021). AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE.

Dueben, P. D., & Bauer, P. (2018, October). Challenges and design choices for global weather and climate models based on machine learning. *Geoscientific Model Development*, *11*(10), 3999–4009. doi: 10.5194/gmd-11-3999-2018

Franch, G., Nerini, D., Pendesini, M., Coviello, L., Jurman, G., & Furlanello, C. (2020, March). Precipitation Nowcasting with Orographic Enhanced Stacked Generalization: Improving Deep Learning Predictions on Extreme Events. *Atmosphere*, *11*(3), 267. doi: 10.3390/atmos11030267

Gao, Y., Guan, J., Zhang, F., Wang, X., & Long, Z. (2022, June). Attention-Unet-Based Near-Real-Time Precipitation Estimation from Fengyun-4A Satellite Imageries. *Remote Sensing*, *14*(12), 2925. doi: 10.3390/rs14122925

Grönquist, P., Yao, C., Ben-Nun, T., Dryden, N., Dueben, P., Li, S., & Hoefler, T. (2021, April). Deep learning for post-processing ensemble weather forecasts. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *379*(2194), 20200092. doi: 10.1098/rsta.2020.0092

Hasan, M. A., & Islam, A. K. M. S. (2018, December). Evaluation of Microphysics and Cumulus Schemes of WRF for Forecasting of Heavy Monsoon Rainfall over the Southeastern Hilly Region of Bangladesh. *Pure and Applied Geophysics*, *175*(12), 4537–4566. doi: 10.1007/s00024-018-1876-z

Hess, P., & Boers, N. (2022, March). Deep Learning for Improving Numerical Weather Prediction of Heavy Rainfall. *Journal of Advances in Modeling Earth Systems*, *14*(3), e2021MS002765. doi: 10.1029/2021MS002765

Jee, J.-B., & Kim, S. (2017, May). Sensitivity Study on High-Resolution WRF Precipitation Forecast for a Heavy Rainfall Event. *Atmosphere*, *8*(6), 96. doi: 10.3390/atmos8060096

Kain, J. S. (2004, January). The Kain–Fritsch Convective Parameterization: An Update. *Journal of Applied Meteorology*, *43*(1), 170–181. doi: 10.1175/1520-0450(2004)043⟨0170:TKCPAU⟩2.0.CO;2

Kain, J. S., & Fritsch, J. M. (1993). Convective Parameterization for Mesoscale Models: The Kain-Fritsch Scheme. In K. A. Emanuel & D. J. Raymond (Eds.), *The Representation of Cumulus Convection in Numerical Models* (pp. 165–170). Boston, MA: American Meteorological Society. doi: 10.1007/978-1-935704-13-3_16

Kalnay, E. (2003). *Atmospheric modeling, data assimilation, and predictability.* New York: Cambridge University Press.

Li, W., Pan, B., Xia, J., & Duan, Q. (2022, February). Convolutional neural network-based statistical post-processing of ensemble precipitation forecasts. *Journal of Hydrology*, *605*, 127301. doi: 10.1016/j.jhydrol.2021.127301

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., … Guo, B. (2021, October). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 9992–10002). Montreal, QC, Canada: IEEE. doi: 10.1109/ICCV48922.2021.00986

Ma, L.-M., & Tan, Z.-M. (2009, April). Improving the behavior of the cumulus parameterization for tropical cyclone prediction: Convection trigger. *Atmospheric Research*, *92*(2), 190–211. doi: 10.1016/j.atmosres.2008.09.022

Otieno, G., Mutemi, J. N., Opijah, F. J., Ogallo, L. A., & Omondi, M. H. (2020, February). The Sensitivity of Rainfall Characteristics to Cumulus Parameterization Schemes from a WRF Model. Part I: A Case Study Over East Africa During Wet Years. *Pure and Applied Geophysics*, *177*(2), 1095–1110. doi: 10.1007/s00024-019-02293-2

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., … Anandkumar, A. (2022, February). *FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators.* arXiv. (arXiv:2202.11214 [physics])

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., & Thuerey, N. (2020, November). WeatherBench: A Benchmark Data Set for Data-Driven Weather Forecasting. *Journal of Advances in Modeling Earth Systems*, *12*(11). doi: 10.1029/2020MS002203

Rasp, S., & Thuerey, N. (2021, February). Data-Driven Medium-Range Weather Prediction With a Resnet Pretrained on Climate Simulations: A New Model for WeatherBench. *Journal of Advances in Modeling Earth Systems*, *13*(2). doi: 10.1029/2020MS002405

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., … Mohamed, S. (2021, September). Skilful precipitation nowcasting using deep generative models of radar. *Nature*, *597*(7878), 672–677. doi: 10.1038/s41586-021-03854-z

Ronneberger, O., Fischer, P., & Brox, T. (2015, May). *U-Net: Convolutional Networks for Biomedical Image Segmentation.* arXiv. (arXiv:1505.04597 [cs])

Salehi, S. S. M., Erdogmus, D., & Gholipour, A. (2017, June). *Tversky loss func-*

tion for image segmentation using 3D fully convolutional deep networks. arXiv. (arXiv:1706.05721 [cs])

Scher, S., & Messori, G. (2019, July). Weather and climate forecasting with neural networks: using general circulation models (GCMs) with different complexity as a study ground. *Geoscientific Model Development*, *12*(7), 2797–2809. doi: 10.5194/gmd-12-2797-2019

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-k., & WOO, W.-c. (2015). Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 28). Curran Associates, Inc.

Shi, X., Gao, Z., Lausen, L., Wang, H., Yeung, D.-Y., Wong, W.-k., & WOO, W.-c. (2017). Deep Learning for Precipitation Nowcasting: A Benchmark and A New Model. In I. Guyon et al. (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc.

Sønderby, C. K., Espeholt, L., Heek, J., Dehghani, M., Oliver, A., Salimans, T., ... Kalchbrenner, N. (2020, March). *MetNet: A Neural Weather Model for Precipitation Forecasting.* arXiv. (arXiv:2003.12140 [physics, stat])

Taillardat, M., & Mestre, O. (2020, May). From research to applications – examples of operational ensemble post-processing in France using machine learning. *Nonlinear Processes in Geophysics*, *27*(2), 329–347. doi: 10.5194/npg-27-329-2020

Tao, Y., Gao, X., Ihler, A., Sorooshian, S., & Hsu, K. (2017, May). Precipitation Identification with Bispectral Satellite Information Using Deep Learning Approaches. *Journal of Hydrometeorology*, *18*(5), 1271–1283. doi: 10.1175/JHM-D-16-0176.1

Tran, Q.-K., & Song, S.-k. (2019, May). Computer Vision in Precipitation Nowcasting: Applying Image Quality Assessment Metrics for Training Deep Neural Networks. *Atmosphere*, *10*(5), 244. doi: 10.3390/atmos10050244

Wang, C., Tang, G., Xiong, W., Ma, Z., & Zhu, S. (2021, December). Infrared precipitation estimation using convolutional neural network for FengYun satellites. *Journal of Hydrology*, *603*, 127113. doi: 10.1016/j.jhydrol.2021.127113

Xie, Pingping, Joyce, Robert, Wu, Shaorong, Yoo, S.-H., Yarosh, Yelena, Sun, Fengying, & Lin, Roger. (2019). *NOAA Climate Data Record (CDR) of CPC Morphing Technique (CMORPH) High Resolution Global Precipitation Estimates, Version 1.NOAA National Centers for Environmental Information.* (NOAA CDR Program (2019)) doi: https://doi.org/10.25921/w9va-q159

Zhou, K., Sun, J., Zheng, Y., & Zhang, Y. (2022, September). Quantitative Precipitation Forecast Experiment Based on Basic NWP Variables Using Deep Learning. *Advances in Atmospheric Sciences*, *39*(9), 1472–1486. doi: 10.1007/s00376-021-1207-7