JGR Atmospheres

RESEARCH ARTICLE
10.1029/2023JD039311

East Asia Atmospheric River Forecast With a Deep Learning Method: GAN-UNet

Yuan Tian[1], Yang Zhao[2,3], Jianping Li[2,4], Bin Chen[5], Lin Deng[6], and Dawei Wen[7]

[1]School of Systems Science, Beijing Normal University, Beijing, China, [2]Frontiers Science Center for Deep Ocean Multispheres and Earth System (FDOMES)/Key Laboratory of Physical Oceanography/Academy of the Future Ocean, Ocean University of China, Qingdao, China, [3]College of Oceanic and Atmospheric Sciences, Ocean University of China, Qingdao, China, [4]Laoshan Laboratory, Qingdao, China, [5]State Key Laboratory of Severe Weather, Chinese Academy of Meteorological Sciences, Beijing, China, [6]Shanghai Typhoon Institute, Shanghai, China, [7]School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan, China

Abstract Accurate forecasting of atmospheric rivers (ARs) holds significance in preventing losses from extreme precipitation. However, traditional numerical weather prediction (NWP) models are computationally expensive and can be limited in accuracy due to inaccurate physical parameter settings. To overcome these limitations, we propose a deep learning (DL) model, called GAN-UNet, to forecast the AR occurrence, position, and intensity in East Asia. GAN-UNet can capture the complex nonlinear relationship between the inputs at the past moment, including the vertically integrated water vapor transport (IVT), zonal wind at 850 hPa (U850), and meridional wind at 850 hPa (V850), and the forecast output (IVT, U850, or V850), whose results are comparable to NWP models. In addition, the average model (AM) by integrating the results generated by GAN-UNet and European Centre for Medium-Range Weather Forecasts (ECMWF) outperforms all the NWP models selected in this study, demonstrating its potential to improve the performance of NWP through the DL method. Specifically, the 5-day average F1 scores of the AM are 0.777 and 0.845, whose values are significantly better than those obtained by ECMWF (0.712 and 0.794) in the two key regions of East Asia; The AM 5-day average intersection over unions are 0.706 and 0.688 while the values of ECMWF are 0.675 and 0.64; in terms of intensity forecast, GAN-UNet and AM exhibited lower differences in most of the intensity bins, except for the final bin with IVT more than 825 kg m^{-1} s^{-1}. With this thorough analysis, GAN-UNet is shown as an effective model to forecast ARs.

Plain Language Summary This work presents a deep learning (DL) model (called GAN-UNet) for the spatio-temporal atmospheric rivers (ARs) forecast. Our goal is to perform long lead-time AR event forecasts accurately and efficiently. The data from 1959 to 2022 over 40°E−180°E, 20°S−60°N are used for training, validation, and testing. The results show that the proposed GAN-UNet can forecast the AR occurrence, position, and intensity accurately and its performance is comparable with the state-of-the-art numerical weather prediction (NWP) model. In addition, the ensemble of GAN-UNet and the NWP model outperform all the NWP models selected in this study. However, The DL model still has certain limitation. Specifically, while GAN-UNet and the ensemble model demonstrate superior forecast capabilities for most intensity bins in AR events compared to NWP models, they are severely underestimated for the grids of the final bin. The results show that the DL model can effectively forecast AR events in East Asia, and further improve the forecast accuracy of NWP by combining it with NWP model, which holds significance in extreme precipitation forecast.

Key Points:

- A deep learning model (GAN-UNet) is developed for efficient and accurate atmospheric river forecast
- GAN-UNet achieves comparable performance to ECMWF and NCEP on atmospheric river forecast
- The outcomes obtained from the average of GAN-UNet and NWP exhibit superior performance compared to all the individual NWP models selected in this study

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

Y. Zhao,
qkylyy520@163.com

Citation:

Tian, Y., Zhao, Y., Li, J., Chen, B., Deng, L., & Wen, D. (2024). East Asia atmospheric river forecast with a deep learning method: GAN-UNet. *Journal of Geophysical Research: Atmospheres*, *129*, e2023JD039311. https://doi.org/10.1029/2023JD039311

Received 24 MAY 2023
Accepted 17 FEB 2024

Author Contributions:

Conceptualization: Yang Zhao
Formal analysis: Yuan Tian, Yang Zhao
Funding acquisition: Jianping Li
Investigation: Yang Zhao
Methodology: Yuan Tian
Project administration: Jianping Li
Software: Yuan Tian
Supervision: Yang Zhao, Jianping Li
Validation: Yang Zhao
Visualization: Yuan Tian, Yang Zhao, Dawei Wen
Writing – original draft: Yuan Tian
Writing – review & editing: Yang Zhao, Jianping Li, Bin Chen, Lin Deng

1. Introduction

Extreme precipitation (EP) is a recurring natural disaster which can have adverse impacts on human lives, economy, and agriculture (Masson-Delmotte et al., 2021; Zhao et al., 2023). Recent studies have shown that EP events are increasing worldwide due to global warming, which is likely to continue in the coming decades (AghaKouchak et al., 2020; Li et al., 2020, 2021; Scoccimarro et al., 2013). Atmospheric rivers (ARs, Zhu & Newell, 1994), moisture passages that carry large amounts of water vapor in the mid-latitudes, have been identified as a critical driver of EP events (Gimeno et al., 2014; Kamae et al., 2021; Lavers et al., 2014, 2020; Payne et al., 2020; Ralph et al., 2006; Waliser & Guan, 2017). Studies have demonstrated that the occurrence of ARs can be used as a reliable predictor of EP as it has a high predictability (Chen et al., 2018; Lavers & Villarini, 2013, Lavers et al., 2014; Lavers & Villarini, 2013).

Based on the important impact of ARs on EP, it is valuable to accurately forecast their occurrence and intensity. Among all the forecast methods, numerical weather prediction (NWP) may be the most accurate and popular currently. Furthermore, the accuracy of NWP models has been steadily improving until now (Alley et al., 2019). However, NWP still has some obvious shortcomings. One of the biggest problems is that the accuracy of NWP is often limited by problems such as insufficient understanding of the interaction of physical processes in the model and systematic bias (Bauer et al., 2015; Geer, 2021). With the rise of deep learning (DL), data-driven DL models have been developed to predict future weather, and they have shown promising results in predicting different sounding meteorological variables (i.e., temperature, specific humidity, geopotential height, and horizontal wind speed) and surface meteorological variables (i.e., 2 m temperature, 10 m wind components, and sea surface pressure) (Bi et al., 2022; Lam et al., 2022; Pathak et al., 2022). Specifically, FourCastNet (Pathak et al., 2022) was the first DL model to demonstrate forecast results comparable to European Centre for Medium-Range Weather Forecasts (ECMWF; Molteni et al., 1996), the existing state-of-the-art NWP forecast archive in the THORPEX (The Observing System Research and Predictability Experiment) Interactive Grand Global Ensemble (TIGGE; Park et al., 2008), but its forecasts did not surpass ECMWF. Pangu weather (Bi et al., 2022), on the other hand, achieved the distinction of surpassing ECMWF. Subsequently, GraphCast (Lam et al., 2022) from Google DeepMind surpassed Pangu weather. These DL models exhibit excellent overall forecasting performance globally, especially Pangu weather and GraphCast, which outperformed ECMWF in most variables. However, the aforementioned models did not specifically focus on AR events, with only FourCastNet providing a rough evaluation of global AR forecasting. As far as we know, there is currently no relevant research utilizing DL models for forecasting AR events directly.

Nevertheless, DL has some applications focused on ARs, including AR detection (Higgins et al., 2023; Prabhat et al., 2021; Tian et al., 2023) and postprocessing of AR forecasting (Chapman et al., 2019, 2022). Tian et al. (2023) employed an ensemble of 20 different DL models to perform semantic segmentation for ARs. Each model in the ensemble was trained independently and the final result was obtained via majority voting. When testing across the whole test dataset, the ensemble of models obtained an IOU for ARs of 40.5% and performed much better than any single model in the ensemble, with none surpassing an IOU of 38.5%. Chapman et al. (2019) developed a convolutional neural network model as a post-process tool to deal with the vertically integrated water vapor transport (IVT) field from NWP. The results indicated that convolutional neural network post-process reduced the root-mean-square error by 9%–17% while increasing the correlation between observations and predictions by 0.5%–12%. Furthermore, Chapman et al. (2022) developed a variety of DL-based probabilistic predictions, which showed that DL post-process methods are computationally cost-effective, easy to implement, and can compete with or outperform the dynamical ensemble's raw model output from the National Centers for Environmental Prediction Global Ensemble Forecast System. Despite our emerging understanding of AR forecasts based on or partly based on DL, whether ARs can be forecasted using solely the DL model, producing comparable results with NWP, remains unknown.

In addition, although there are some studies on AR forecasts, most of them are concentrated in North America (Chapman et al., 2019, 2022; DeFlorio et al., 2018; Nardi et al., 2018; Nayak et al., 2014; Wick et al., 2013). As a region where ARs also frequently occur, similar research is lacking in East Asia (Arabzadeh et al., 2020; Espinoza et al., 2018; Kim et al., 2021; Wang et al., 2021). In East Asia, recent studies primarily focused on AR detection (Liang & Yong, 2021; Pan & Lu, 2019; Tian et al., 2023) and the impact of AR-induced EP and flooding (Kamae et al., 2021; Liang et al., 2022; Liang & Yong, 2021). Given the tight link between ARs and EP in East Asia (Liang et al., 2022; Liang & Yong, 2021; Wang et al., 2021), it is meaningful to forecast the AR events in this region.

In light of the aforementioned research gaps and limitations, we develop a DL model called GAN-UNet, which is based on Generative Adversarial Networks (GANs; Goodfellow et al., 2020), to forecast the spatio-temporal ARs over East Asia, evaluate GAN-UNet performance using ERA5 reanalysis data as benchmarks, and compare it with the NWP models. Specifically, we use stacked components in GAN-UNet to describe the details of the forecasts more accurately and implement a hierarchical temporal aggregation strategy to reduce the number of iterations and thus avoid the error accumulation. Overall, the purpose of this study is three-fold: First, to design a DL forecast model for efficient and accurate AR forecast; Second, to verify the feasibility of the DL model by comparing the results of GAN-UNet with those obtained from multiple NWP models; Finally, to integrate GAN-UNet with the state-of-the-art NWP model to explore whether the DL model can improve the results of NWP model.

Table 1
Categories of Precipitation Days

Precipitation levels	Definition
Dry moments	$\leq$0.1 mm day^{-1}
Wet moments	>0.1 mm day^{-1}
EP moments	>90th percentile in the wet moments
Mid-precipitation (MP) moments	wet moments after excluding EP moments

The remainder of this paper is organized as follows: Section 2 describes data and methods. Section 3 first illustrates the importance of ARs to EP and then demonstrates the forecasting effectiveness of GAN-UNet for AR events compared to NWPs. The quality of forecasting effectiveness is manifested in the accuracy of the forecasted AR events occurring in the key regions, the difference of the AR events in profile and intensity between the forecasts and the labels. Conclusions and discussion are provided in Section 4.

2. Data and Methods

2.1. Data

The data used in this study as input variables and labels for GAN-UNet are the fifth generation of ECWMF atmospheric reanalysis data (ERA5, Hersbach et al., 2020) from 1959 to 2022, including specific humidity (q), zonal wind (u), meridional wind (v), and precipitation for the warm season from May to August. The time resolution of q, u, and v is 6 hr (0000/0600/1200/1800 UTC), and precipitation is at hourly resolution.

In 2005, a World Weather Research Program called TIGGE was launched at a workshop at ECMWF, for which one of the main objectives was to strengthen cooperation between operational centres and universities in the development of ensemble forecasting (Swinbank et al., 2016). The TIGGE archive includes forecast data from 10 different operational weather centers, including the ECMWF, the Japan Meteorological Agency (JMA), the Korean Meteorological Administration (KMA), the US National Centers for Environmental Prediction (NCEP) and so on. To compare the forecast effects with GAN-UNet, control forecast data provided by TIGGE are used. The products that have been widely confirmed to exhibit excellent forecast performance with lead times of up to 15 days are selected (Bi et al., 2022; Lam et al., 2022; Nardi et al., 2018; Wick et al., 2013). Specifically, the products are selected from (a) the Environment and Climate Change Canada (ECCC), (b) ECMWF, and (c) NCEP. ECCC and ECMWF only provide the forecasts for 0000 and 1200 UTC. All the models provide forecasts up to a lead time of 360 hr with 6-hr steps. To achieve optimal time consistency with the available data in the NWP, the comparison study uses 0000 and 1200 UTC as starting times, 6-hr steps, and a lead time of 360 hr for both the DL and NWP models. All the variables mentioned above apply with a spatial resolution of 1.5° × 1.5°.

2.2. Definition of Different Precipitation Levels, AR-Related Precipitation, and AR Events

In East Asia, rich water vapor from the low latitude is transported during the summer monsoon season, which brings notable precipitation to eastern China, the Korean Peninsula, and Japan. We focus on two key study regions: East China (EC, 27°–34.5°N, 114°–123°E) and Korea and Japan (KJ, 31.5°–39°N, 126°–135°E), land areas where ARs occur frequently. In this study, we divide all precipitation days into the following categories based on precipitation intensity (Table 1). The AR frequency is the number of AR occurrences divided by all the time steps. The AR-related precipitation is defined as the precipitation covered by the AR grid (Prabhat et al., 2021; Tian et al., 2023; Zhang et al., 2023). When an AR occurs and occupies more than 50% areas of a key region, we define it as an AR event. An AR event intensity is defined as the average IVT intensity of the AR event.

2.3. Methods

2.3.1. AR Detection Algorithm

ARs are defined as abnormally enhanced plume water vapor transport. We use the AR detection algorithm similar to previous studies (Guan & Waliser, 2015; Liang & Yong, 2021; Mundhenk et al., 2016), but with some simplifications. The ARs over 40°E−180°E, 20°S-60°N (94 × 54 grids) are detected by the following steps:

(1) Calculating the IVT using the reanalysis and forecast data. The IVT is defined as:

$$\mathrm{IVT} = \sqrt{\left(-\frac{1}{g}\int_{ps}^{pt} qudp\right)^2 + \left(-\frac{1}{g}\int_{ps}^{pt} qvdp\right)^2} \quad (1)$$

where g is gravitational acceleration, p is atmospheric pressure in p-coordinate, ps is the surface pressure, pt is the atmospheric top layer pressure, and here is 300 hPa.

(2) Scanning the grids of IVT greater than 500 kg m^{-1} s^{-1}. In this study, an instantaneous absolute IVT threshold of 500 kg m^{-1} s^{-1}, rather than a relative threshold like Guan and Waliser (2015) and Pan and Lu (2019), is preferred. It is because the forecast products of TIGGE start from 2008, it is not possible to calculate the relative threshold for consistency with the reanalysis data, which leads to uncertainty in the results of different models. Moreover, 500 kg m^{-1} s^{-1} is considered to be the lower bound for moderate to strong ARs and accounts for a significant portion of the AR-related hazards (Liang & Yong, 2021; Mahoney et al., 2016; Nardi et al., 2018; Reid et al., 2020).

(3) Detection and isolation of continuous regions of the identified grids, and calculation of the axis length and width of each isolated region. The axis length of each isolated region must be greater than 2,000 km, and the aspect ratio greater than 2 (Guan & Waliser, 2015).

(4) If the above criteria are met, the coverage of the isolated region is defined as an AR.

2.3.2. The Proposed Model Architectures and Related Parameters

GANs (Goodfellow et al., 2020) have been successful in a wide range of applications, including weather forecast (Gong et al., 2022; Ravuri et al., 2021). The basic idea behind GANs is to train two neural networks simultaneously in a two-player minimax game framework, where the generator tries to generate realistic samples that can fool the discriminator, and the discriminator tries to distinguish between the real and fake samples generated by the generator. In weather forecast, GANs can be used to generate high-quality weather forecasts by training the generator on historical weather data and using it to generate future weather predictions. The discriminator can then be used to evaluate the accuracy of the generated forecasts by comparing them with actual weather data.

However, a single GAN may not always be adequate for generating high-quality outputs that accurately represent the underlying distribution, particularly when the target distribution is highly complex and contains intricate details that are challenging to capture using a single generator. To address this limitation, researchers have proposed the use of multiple GANs that are arranged in a hierarchical fashion, referred to as "stacked GANs" or "hierarchical GANs" (Durugkar et al., 2016; Ghosh et al., 2018; Karras et al., 2017).

Drawing inspiration from stacked GANs, the present study proposes a novel GAN architecture referred to as GAN-UNet, comprising of two generators and one discriminator, in order to effectively capture the fine details of the forecast results. Datasets from 1959 to 2022 over 40°E−180°E, 20°S−60°N are used for training, validation, and testing. Similar to Bi et al. (2022), one year (2003) is used for validation, three years (2008, 2013, and 2018) for testing, and the rest for training. However, unlike Bi et al. (2022), this study uses a discontinuous selection strategy, which can avoid different climate change signals due to different periods of the training and testing datasets (Chen & Wang, 2022; Tian et al., 2023). The proposed GAN-UNet framework is composed of four main components (Steps 1–4), as illustrated in Figure 1. In Step 1, the generator1 is employed to generate a rough sketch of the forecasted IVTs, producing results 1. In Step 2, the discriminator is trained using gradients computed from the loss function passed to it from generator1. During training, generator1 and the discriminator play a min-max game, where generator1 aims to minimize the loss while the discriminator strives to maximize it. This min-max game is primarily reflected in the loss function, where, with each iteration of training, the sum of the losses between the results generated by generator1 and the discriminator and their respective labels becomes increasingly smaller. The specific form of the loss function is described in Equation 3 in the following text. As the training progresses, both the generator1 and discriminator become more adept at their respective tasks until the discriminator can no longer distinguish between the results 1 generated by generator1 and the actual labels. Using the trained generator1 and discriminator, we obtain results 2. In Step 3, we refine the results two obtained in Step 2 by utilizing generator2 for training. Finally, as the frequency of extreme values, including maximum and minimum values, is significantly lower in space and time, the trained network tends to underestimate its maximum while overestimating its minimum (Bi et al., 2022; Chen & Wang, 2022). To address this issue, we fine-tune the results three obtained in Step 3 using a univariate cubic equation, thereby enhancing the network's performance in the extreme IVT in Step 4:

$$\mathrm{IVT}_t(i) = a(\mathrm{IVT}_r(i))^3 + b(\mathrm{IVT}_r(i))^2 + c\mathrm{IVT}_r(i) + d \tag{2}$$

where i denotes the longitude-latitude grid index; $\mathrm{IVT}_r(i)$ and $\mathrm{IVT}_t(i)$ are the results three (i.e., without tuning) and tuned IVT forecast value at the ith grid. The parameters a, b, c, and d are four tunable parameters, which are

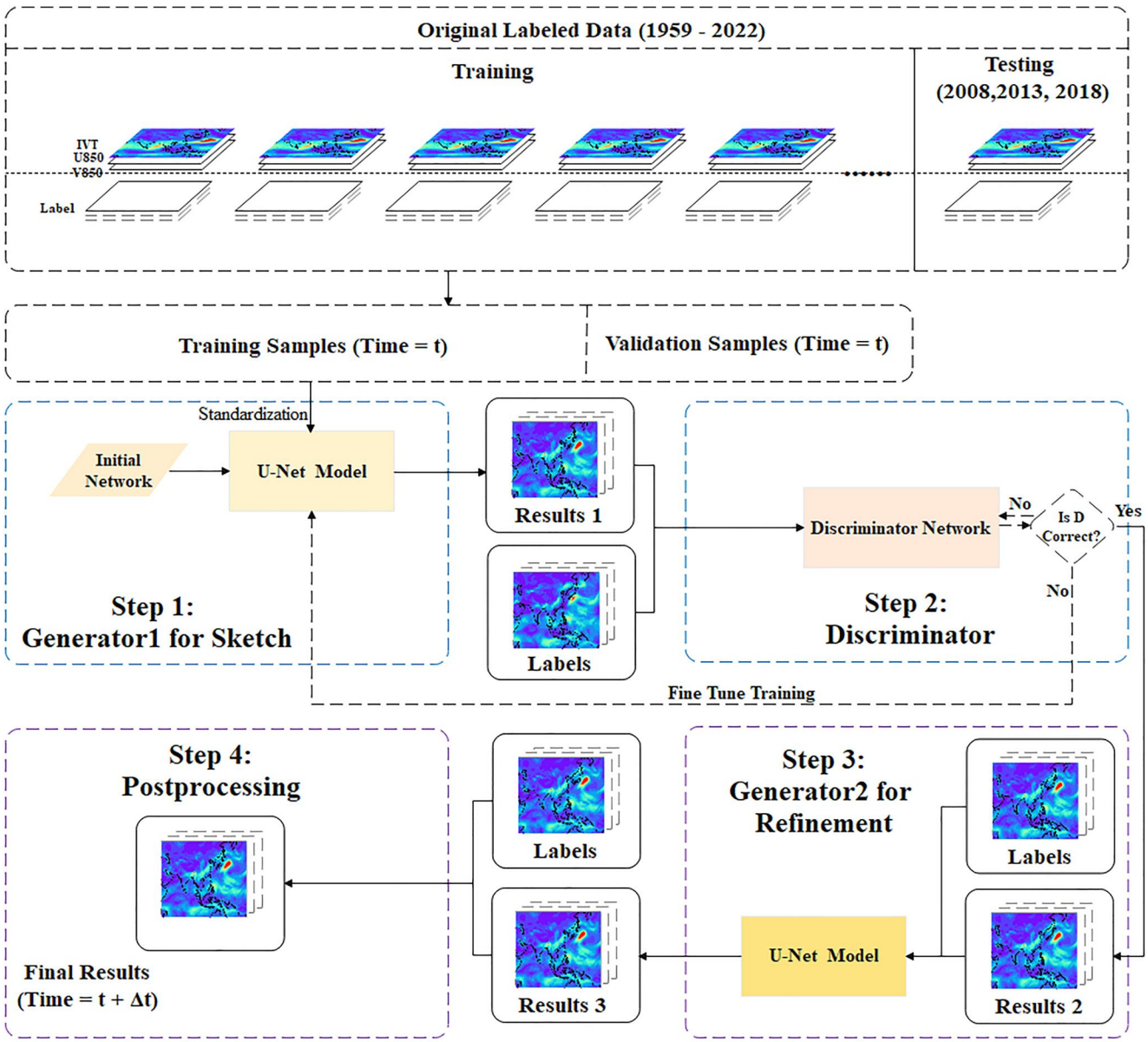

Figure 1. The overall forecast process of GAN-UNet model. GAN-UNet contains three input and output layers (i.e., IVT, U850, and V850). Since the inputs of a single model is all the three variables and the output is one variable, all three variables must be forecasted separately by running the model for each variable.

determined by the training data output after model steps 1–3 and the ERA5 reanalysis data for the corresponding true moment (i.e., the label).

The components of GAN-UNet are briefly outlined as follows. The generators including generator1 and generator2 used in the model are both the U-Net model (Ronneberger et al., 2015). Similar as Tian et al. (2023), the inputs to the generator1 are composed of the datasets with three variables, namely IVT, zonal wind at 850 hPa (U850), and meridional wind at 850 hPa (V850), which mainly consider that the wind fields at 850 hPa can offer a reliable representative measure of the AR winds as 850 hPa corresponds closely to the central altitude at which the low-level jet stream is typically situated. At the same time, the additional atmospheric variables at other levels in the atmosphere (e.g., 500-hPa) may not show any significant improvement because the information from these upper-level fields (e.g., jet variability) is indirectly captured in the lower-level atmospheric fields to a sufficient degree (). Before entering the model, each variable is first standardized to: (a) eliminate the influence of different variable scales; (b) reduce the impact of extreme values; (c) accelerate the speed of gradient descent for optimal

solution. The results 1–3 and the final results are composed of the datasets with only one variable (IVT, U850, or V850) for future time steps. It is noteworthy that the inputs of GAN-UNet are the data at time t, while the results represent the forecasted variable for time t+1, t+2, or t+3. As noted by Bi et al. (2022), we train three individual models for 6-hr, 12-hr, and 18-hr forecast, respectively and iterate to generate the forecasts for subsequent moments, which is called as hierarchical temporal aggregation. For example, when we need to forecast AR events 6 hr ahead, there is no doubt that we should use the model designed for the 6-hr forecast. However, when forecasting AR events 24 hr ahead, we first forecast the state 6 hr ahead and then input this result into the model designed for the 18-hr forecast. When forecasting AR events 48 hr ahead, we input the data twice into the model designed for the 18-hr forecast and then input it into the model designed for the 12-hr forecast. The discriminator used in the model is a variation of the traditional convolutional neural network model.

GAN-UNet model employs a multi-scale loss function that extracts hierarchical features from multiple layers of the discriminator. This loss function captures both long-range and short-range spatial relationships between pixels by utilizing features at different levels of granularity, including pixel-level features, low-level features (such as superpixels), and medium-level features (such as patches). The loss function V used in GAN-UNet model can be defined as follows:

$$\min_G \max_D V(D,G) = E_{x\sim P_{data}(x)}[\log D(x)] + E_{Z\sim P_z(z)}[\log(1 - D(G(z)))] \tag{3}$$

G and D represent the generator1 and the discriminator, respectively. E represents the mathematical expectation. x represents the labels, and z represents the input variables of the generator1. $G(z)$ means the results 1 generated by the generator1. $D(G(z))$ represents the probability that the input of the discriminator is determined to be the labels. The generator1 aims to maximize the value of $D(G(z))$, while the discriminator aims to maximize $D(x)$ and minimize $D(G(z))$. Therefore, the objective is to minimize $V(D,G)$ by optimizing the generator1 ($\min_G$) and maximizing $V(D,G)$ by optimizing the discriminator ($\max_D$).

The mean squared error (MSE) is utilized as the loss function after the generator2, which is calculated as of the difference between the forecasts and the labels. The equation of MSE is as followed:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \tag{4}$$

where i denotes the longitude-latitude grid index, $\widehat{Y}_i$ is the label, and Y_i is the forecast.

In addition, to assess whether DL can improve the forecast of NWPs, GAN-UNet is combined with the ECMWF forecast via averaging with equal weights, which is same as Chen and Wang (2022). The combined forecast can be expressed as:

$$\mathrm{AM}_i = 0.5\mathrm{GAN\text{-}UNet}_i + 0.5\mathrm{ECMWF}_i \tag{5}$$

where i denotes the longitude-latitude grid index, AM_i is the average model (AM).

2.3.3. Performance Evaluation

The evaluation metrics used in this study are focused on assessing the accuracy of GAN-UNet in forecasting the location and intensity of the AR events. The evaluation is done from three different aspects. (a) Classification performance: This aspect evaluates whether an AR event occurs in the key regions. This is a dichotomous problem, with only two outcomes: positive (an AR event is present) or negative (an AR event is absent). We use the confusion matrix to represent the classification results (Figure 2). Hence, based on whether the AR event of the forecast or the label occurs, there are four situations at last:

① Hit: the AR event occurrence is forecasted correctly.
② False alarm: the model forecast an AR event, but the label does not.
③ Miss: the model does not forecast an AR event, but the label does.
④ Correct reject: the model correctly forecast an absent AR event.

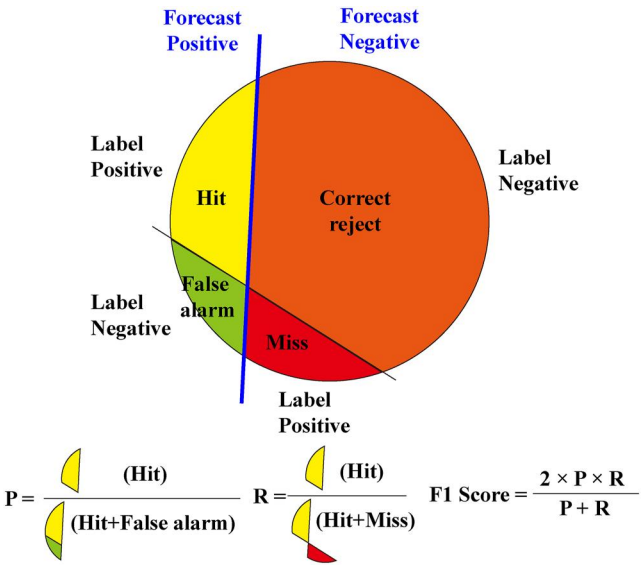

Figure 2. Confusion matrix of the classification results and the expressions of P, R, and F1 score.

Three specific indicators, namely precision (P), recall (R), and F1 score are calculated based on these four outcomes. P is for the forecasts, which means the probability of being forecasted correctly to occur in the samples where the AR events are forecasted to occur. That is, P = number of correctly forecasted AR events/total number of events that are forecasted as ARs. R is for the labels, which means the probability of being forecasted correctly to occur in the samples where the AR events actually occur. R = number of correctly forecasted AR events/total number of the AR events. F1 score is defined as a harmonic average of P and R, which helps to find a balance between them. The formulas for the three indicators are shown in Figure 2; (b) Intersection over union (IOU; Tian et al., 2023): This aspect measures the degree of the AR event overlap for the forecasts and the labels in the key regions. IOU is used as the evaluation metric for this aspect, which is defined as the ratio of the intersection area between the forecast and the label to the union area of these two; (c) Intensity difference: This aspect focuses on the differences in intensity between coincident AR events of the forecasts and the labels. This evaluation is only done for the hit outcomes from the classification performance evaluation.

3. Results

3.1. The Importance of ARs to EP

This section investigates the differential impacts of ARs in MP and EP moments within East Asia, as well as the ratios of the AR events across different precipitation bins. The analysis is based on Figures 3a and 3b, which reveal that AR frequency and intensity are highest in the two terrestrial key regions within East Asia. Specifically, AR frequency is considerably higher in EP moments compared to MP moments, particularly in these two key regions. This finding aligns with previous studies (Guan et al., 2023; Kim et al., 2021; Liang & Yong, 2021; Ralph et al., 2006; Slinskey et al., 2020; Waliser & Guan, 2017). Similarly, the AR intensity is also stronger in EP moments relative to MP moments. These results confirm the close association between AR occurrence and EP.

Furthermore, Figures 3c and 3d demonstrate that the AR-related precipitation ratio in EP moments are significantly higher than in MP moments in the two key regions. Specifically, for MP moments, the area-averaged ratios of AR-precipitation are 22.5% and 39.8% in EC and KJ, respectively. However, for EP moments, ARs are associated with 40.5% and 60.6% of total precipitation. It underscores the greater impact of ARs to precipitation in EP moments relative to MP moments within East Asia.

To further support the notion that ARs are associated with EP, the study investigates the ratios of AR events across different precipitation bins (Figure 3e). The total numbers of AR events occurring in EC and KJ are 2,441 and 3,370 between years 1959 and 2022, respectively (not shown in the figure). With the increasing of the precipitation intensity, the ratios of the AR events increase. Overall, the ratios are 1.4% in EC and 0.1% in KJ in dry moments, 7.4% and 7.5% in MP moments (which are calculated as the mean values of the ratios from <10% to 80%–90%), while 32.2% and 31.9% in EP moments, respectively.

In summary, this section provides insights into the impacts of ARs in MP and EP moments within East Asia, and highlights the important impact of ARs in EP moments in EC and KJ. It implies that effective AR forecast is meaningful for EP. Therefore, the subsequent sections will assess the forecast performance of GAN-UNet and AM and compare it with the outcomes of NWPs to establish their efficacy.

3.2. Forecast AR Event Occurrence

The remaining analyses target the questions related to the occurrence of the AR events, location and intensity difference between the forecasts and the labels, all based on the data from 2008, 2013, and 2018 (the testing datasets as shown in Figure 1). Since there are totally 60 forecast steps, in order to make the presentation of the results more concise, the forecast results are regridded as a daily temporal resolution.

The first question addressed is to access the frequency of correctly forecasting an AR event when one actually does occur. Figure 4 shows the scoring results of R, P, and F1 score. Overall, GAN-UNet and AM forecast the

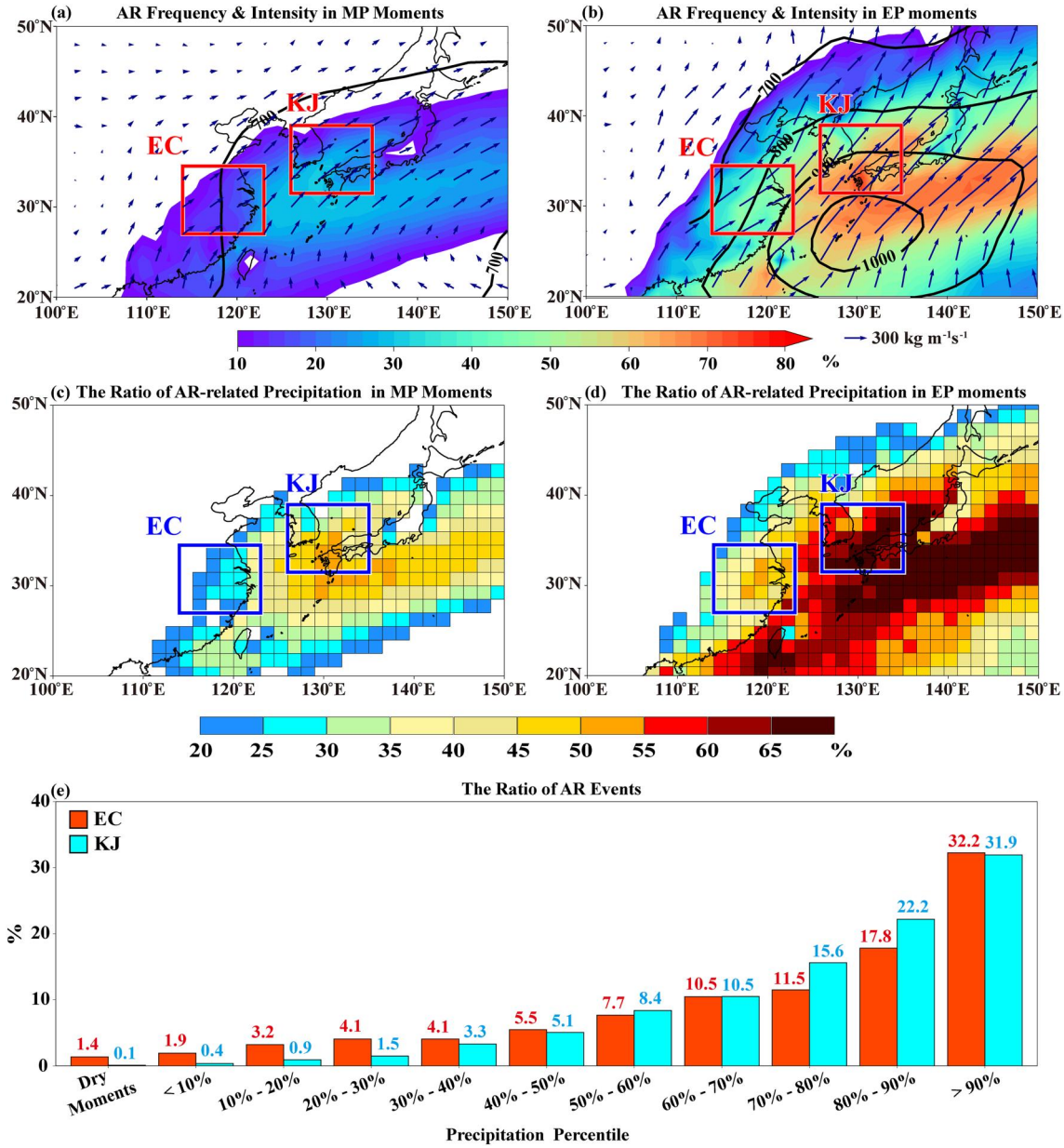

Figure 3. Seasonal-mean (MJJA) climatology of AR frequency (%, shaded), IVT (kg m^{-1} s^{-1}, contoured), and $\overrightarrow{\mathbf{IVT}}$ (kg m^{-1} s^{-1}, vectors) in (a) MP moments and (b) EP moments; (c)–(d) are same as (a)–(b), but for the ratio (%, shaded) of AR-related precipitation to total precipitation; (e) The ratio of the AR events in different precipitation bins in EC and KJ. The reference domains for EC and KJ are boxed in (a)–(d).

likelihood of the AR event occurrence well (basically greater than 0.5) within 5 days, and the forecast effect gradually deteriorates over lead times, similar to the NWPs. For the sake of convenience in description, we refer to the lead times of the first 5 days as "the early lead times" and the lead times after 5 days as "the late lead times". In East Asia, at the early lead times, the forecast performance of AM is the best, the score of the GAN-UNet is slightly lower than ECMWF, but better than NCEP and ECCC. However, the R, P, and F1 score of all the models in this study exhibit inadequate performance in both EC and KJ at the late lead times. The score detail in both two key regions is shown following.

In EC, AM always performs best at the early lead times, followed by GAN-UNet and ECMWF. The 5-day average F1 scores at the early lead times for the 5 models from high to low are 0.755 (AM), 0.712 (ECMWF), 0.712 (GAN-UNet), 0.7 (NCEP), and 0.632 (ECCC). By analyzing the spatial images (not shown),

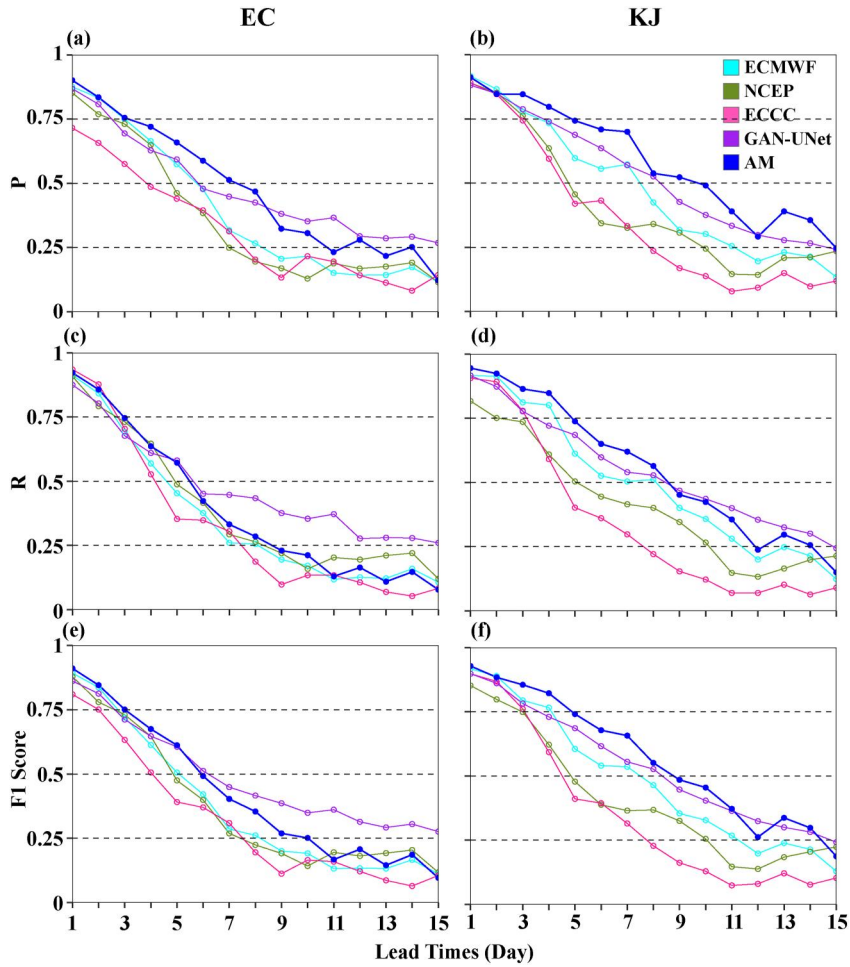

Figure 4. Occurrence-based (a),(b) R, (c),(d) P, and (e),(f) F1 score performances in EC and KJ as a function of forecast lead times.

it is observed that the AR events incorrectly forecasted by ECMWF and GAN-UNet tend to differ at the early lead times and this difference is attributable to minor deviations in IVT. AM mitigates the deviations by computing their mean values, which enhances the overall forecast accuracy. At the late lead times, both AM and GAN-UNet perform notably better than NWPs, but the performance of AM is not as good as GAN-UNet. We think that the reason why GAN-UNet significantly performs better at the late lead times is mainly due to the use of the hierarchical temporal aggregation strategy, which effectively avoids the accumulation of errors in the iterative forecasting process (Bi et al., 2022). Meanwhile, As the average of the GAN-UNet and ECMWF forecasts, the performance of AM falls between the two. As shown in Figure 4e, the inferior performance of AM compared to GAN-UNet could be attributed to the F1 score of ECMWF falling below 0.5, indicating that ECMWF's forecast may have a notable error. At the same time, while GAN-UNet performs best after 5 days, its F1 score is still below 0.5, suggesting that after 5 days, none of the models could maintain accuracy in EC. While in KJ, all the models perform better than in EC within all the lead times. AM performs best at the early lead times, and its performance is on par with GAN-UNet at the late lead times. However, both outperform the NWP models selected in this study. This may be due to the fact that even if ECMWF judgment is incorrect, the IVT differences are still within an acceptable range, so the difference can be reduced by averaging with GAN-UNet within all the lead times. The 5-day average F1 scores at the early lead times for the 5 models from high to low are 0.845 (AM), 0.794 (ECMWF), 0.791 (GAN-UNet), 0.71 (ECCC), and 0.698 (NCEP). With the lead times progress, the R, P, and F1 scores decrease in KJ.

Overall, it indicates that GAN-UNet and AM can accurately forecast the occurrence of AR events in the early lead times, while AM performs best among all the 5 models in this study. Compared with the results in the early

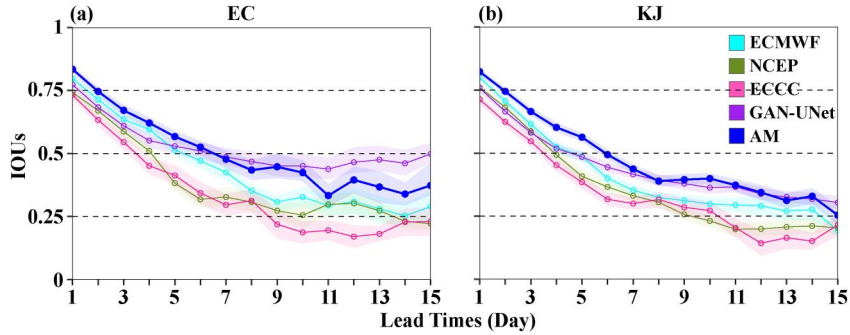

Figure 5. Mean IOUs at (a) EC and (b) KJ with lead times for all the selected models. Shading indicates the 95% confidence intervals of the IOUs.

lead times, all the 5 models cannot maintain a high accuracy on forecasting the AR event occurrence in the late lead times.

3.3. Forecast AR Event Position

To further illustrate the network forecast capability, we second consider the issue of the overlap of the AR events between the forecasts and the labels occurring in the key regions. Recalling the occurrence-based outcomes, a model may forecast an AR event correctly for a moment, but its location may differ from the label. To better quantify the error in the location of the AR events, IOU is used to evaluate the degree to which the forecasts and the labels coincide with those AR events that are correctly forecasted.

As presented in Figure 5, at the early lead time, the 5-day average IOUs of the 5 models from high to low are 0.706 (AM), 0.675 (ECMWF), 0.645 (GAN-UNet), 0.603 (NCEP), and 0.555 (ECCC) in EC; and 0.688 (AM), 0.640 (ECMWF), 0.613 (GAN-UNet), 0.606 (NCEP), and 0.544 (ECCC) in KJ. The results indicate that AM outperforms the other models, with ECMWF and GAN-UNet following closely behind, while NCEP and ECCC exhibit the relatively poor performance at the early lead time. Similar to the occurrence of the AR events, as the lead times progress, none of the models can maintain high IOUs, most of which are below 0.5.

3.4. Forecast AR Event Intensity

The third crucial aspect of effective AR event forecast is the correct forecast of the AR event intensity. Even if a model can accurately forecast the AR event location, a large error in the intensity may still occur. Figure 6 evaluates the AR event intensity difference between the forecasts and the labels for various lead times. The radar chart facilitates the assessment of the forecast skill for individual lead time, with a polygon's perimeter closer to 0 (black bold curve) indicating better performance.

To quantify the intensity difference between the forecasts and the labels, while avoiding the effect of the average difference over multiple lead times that may lead to a reduced average difference of positive and negative values, the mean absolute value of the intensity difference at the early (late) lead times is calculated. For example, assuming during the lead times of 1–5 days, the intensity differences of ECMWF are 0, 15, −15, 10, and 10, then its average intensity difference is 0. However, its mean absolute value is 10. The mean absolute values of the intensity difference of the 5 models at the early lead times are respectively 12.219 (ECMWF), 3.612 (NCEP), 25.011 (ECCC), 36.373 (GAN-UNet), and 39.645 (AM) in EC; and 11.944 (ECMWF), 13.088 (NCEP), 11.348 (ECCC), 48.473 (GAN-UNet), and 49.682 (AM) in KJ. While at the late lead times, the values are 27.586 (ECMWF), 13.836 (NCEP), 44.828 (ECCC), 58.962 (GAN-UNet), and 100.417 (AM) in EC; and 16.609 (ECMWF), 20.269 (NCEP), 18.785 (ECCC), 65.842 (GAN-UNet), and 99.650 (AM) in KJ. It is evident that GAN-UNet and AM show poor forecast results than the other models, with a severe underestimation of the labels' intensity. With the lead times going on, AM forecasts become increasingly severely underestimated, making it the worst-performing model. As mentioned earlier, errors in the models are likely to be larger at extreme values, including the maximum and minimum. However, since in the process of AR detection, the low values of IVT are removed (only consider IVT >500 kg m^{-1} s^{-1}), the largest IVT values may have the most significant impact on the AR event forecast. In other words, a smaller forecast of the maximum values is an important reason for the

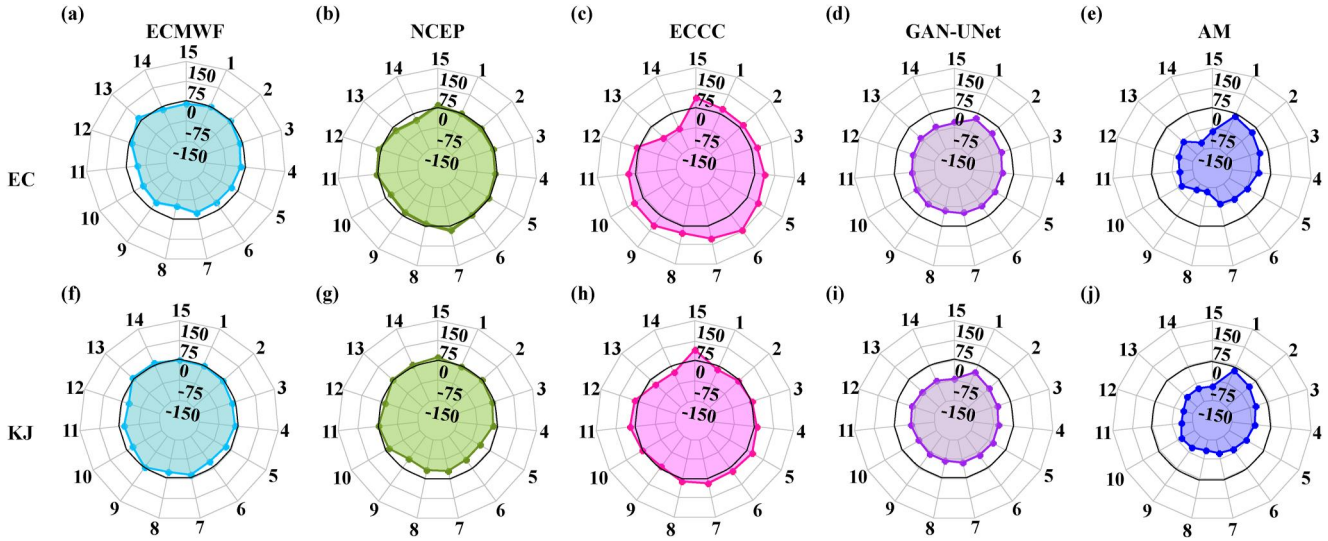

Figure 6. The radar charts of the AR intensity difference (kg m^{-1} s^{-1}) between forecasts based on (a), (f) ECMWF, (b), (g) NCEP, (c), (h) ECCC, (d), (i) GAN-UNet, and (e), (j) AM and the labels over EC and KJ. The numbers outside the circles represent lead times (day). The numbers inside the circles represent the value of AR intensity difference corresponding to each circle. The black bold curve represents a value of AR intensity difference of zero.

smaller AR event intensity forecast. To test this hypothesis, we divide all pixels that correctly forecasted AR events into different intensity bins to compare the difference between the forecasts and the labels, as shown in Figures 7 and 8.

To investigate the errors of AR event intensity across varying forecast lead times, the initial, middle, and final lead times, specifically representing 1-, 7-, and 15-day are chosen, respectively. As shown in Figure 7, at 1-day lead time, the difference between the forecasts and the labels across different IVT bins is relatively small, especially in the case of ECMWF and AM. The box plot length of AM is smaller than that of ECMWF, indicating that the forecasts of all the grids in AM is closer to the labels. However, in the maximum IVT bin, the forecast value of AM is low, which may be attributable to GAN-UNet's underestimation of the final bin. This finding helps to explain why, at the early lead times, the average IVT forecast values for GAN-UNet and AM are lower than the labels (Figure 6). NCEP forecasts are higher for small IVT bins, and lower for large bins, accounting for its small total IVT differences. Conversely, ECCC forecasts for all bins are generally higher, consistent with the larger forecasts reported in Figure 6.

At 7- or 15-day lead time, all the 5 models show larger differences across all bins, including both mean values and standard deviation. For ECMWF, the mean values of the forecasts are higher for small IVT bins, and lower for large bins, resulting in small total IVT differences (Figures 6a and 6f). The forecasts for GAN-UNet and AM models in different bins are similar to those of the ECMWF model, but the standard deviation is smaller and the underestimation of the largest bin is more pronounced. This could be the reason for their poor performance in Figure 6. Moreover, in the minimum IVT bin, AM has the lowest difference, perhaps explaining why it is the best at forecasting AR event occurrence and position. In addition, the length of the box plot of AM is also the smallest among all bins, indicating the minimal difference fluctuation. For NCEP and ECCC, the forecasts are higher than or equal to the labels for all bins, and the differences of the mean of the two decreases with increasing IVT intensity, except for the largest bin where the forecasts are lower than the labels. The models in KJ behave similarly to those in EC, except that the box plot lengths are larger, indicating greater variability in forecast-label differences across bins.

To further verify the possibility that GAN-UNet and AM's poor performances in Figure 6 is mainly due to its small forecasts in the final bin, we conduct a sensitivity experiment by replacing all the forecast and label grids with 825 kg m^{-1} s^{-1} when IVT > 825 kg m^{-1} s^{-1} at the corresponding grids in the labels. The results in Figure 9 show that GAN-UNet forecast is slightly larger, and the errors of AM is minimal, the forecasts of ECMWF, NCEP, and ECCC are all on the large side, which confirms the results in Figures 7 and 8. Based on the preceding analysis, it can be inferred that the conclusion that even though GAN-UNet and AM forecast the AR event

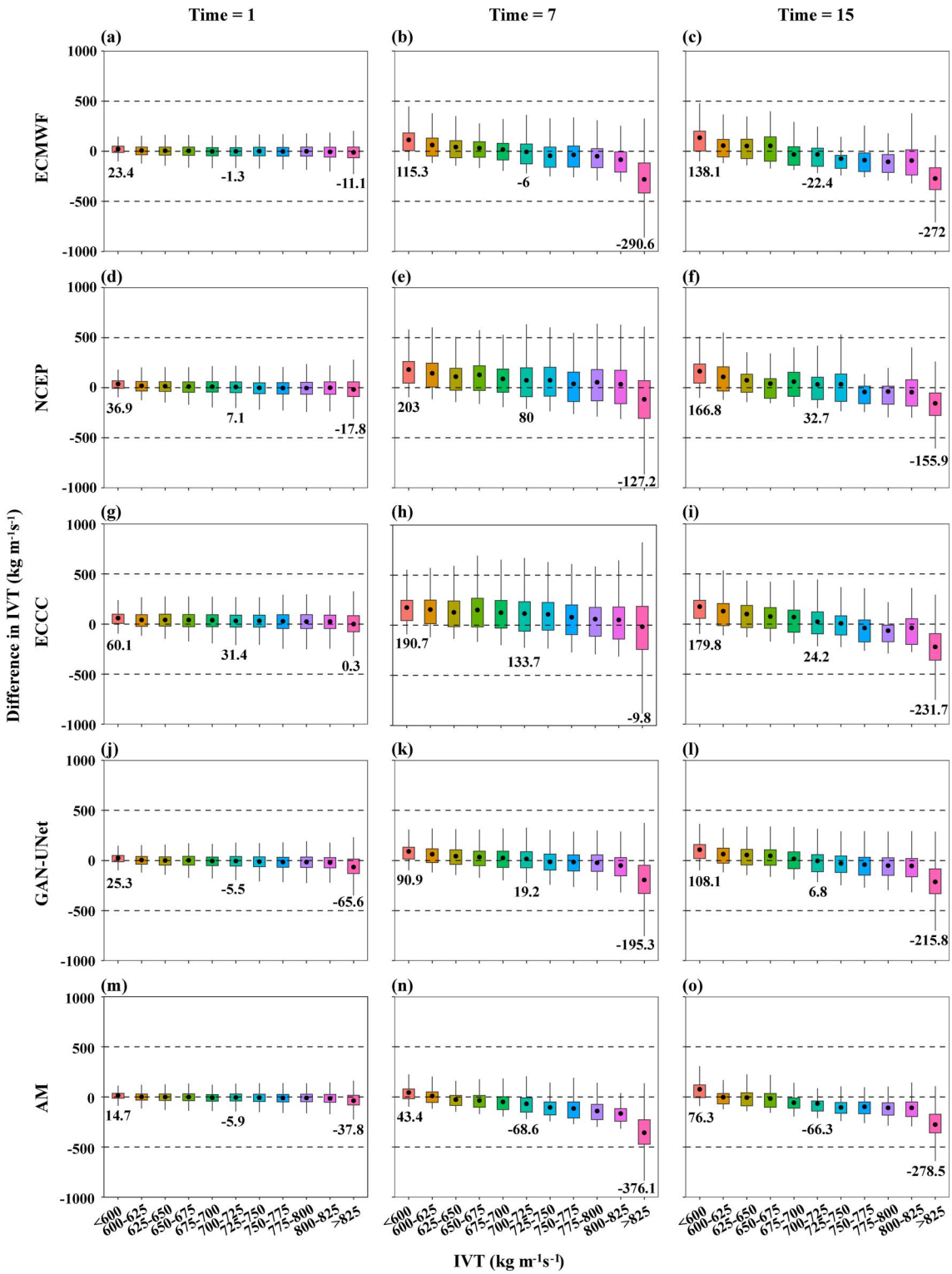

Figure 7. Box-and-whisker plots binned by IVT (kg m^{-1} s^{-1}) of (a–c) ECMWF, (d–f) NCEP, (g–i) ECCC, (j–l) GAN-UNet, (m–o) AM at lead time = 1, 7, and 15 days in EC. The box covers the interquartile range, the black circle inside the box is the mean values, and the whiskers indicate the range (i.e., less than 1.5 times the interquartile range above or below the upper or lower quartile, respectively). The mean values of the initial, middle, and final bins are represented by black values.

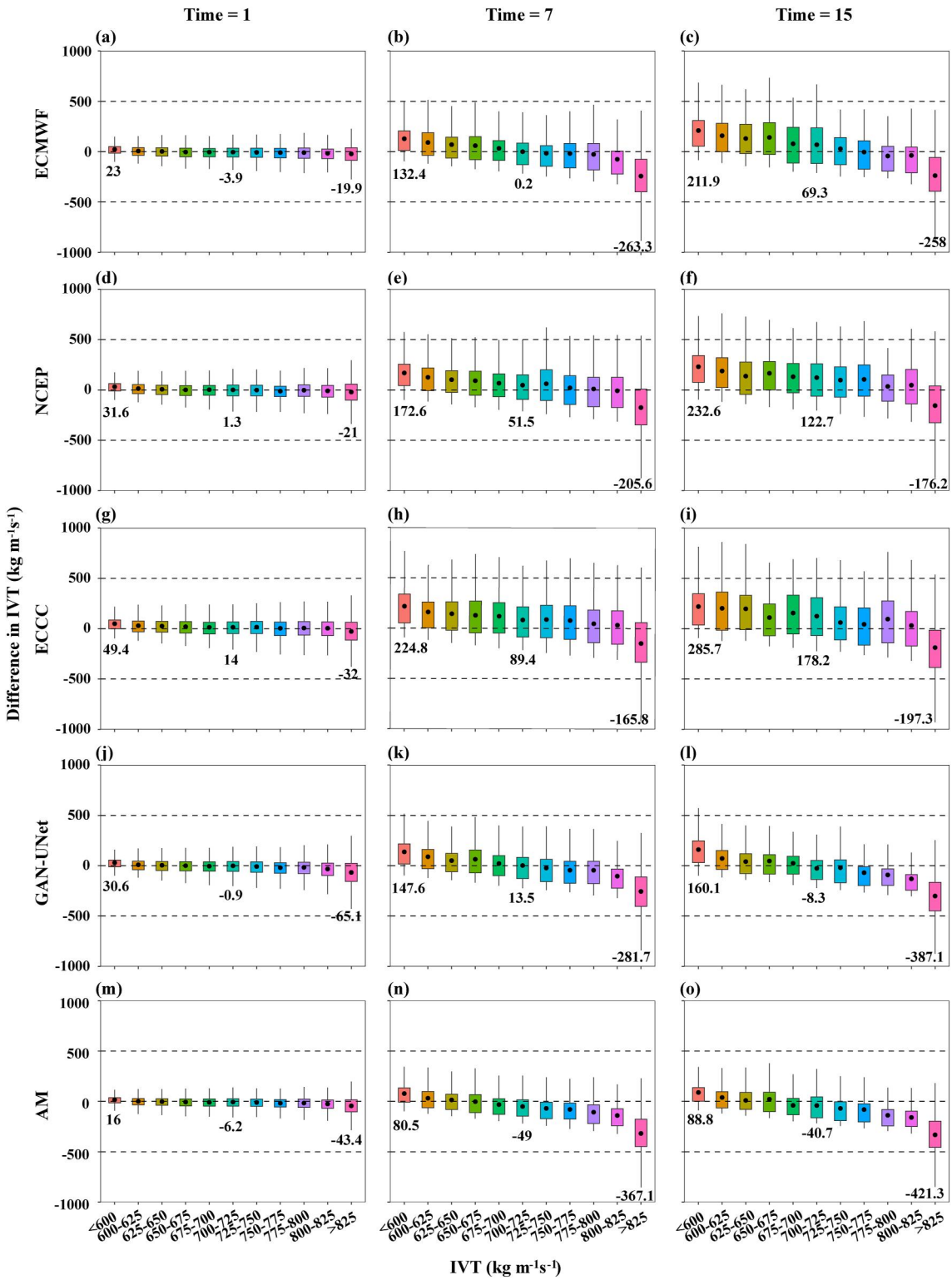

Figure 8. As in Figure 7 but in KJ.

occurrence and location accurately, its forecasts of the maximum values are severely underestimated. In summary, as lead times advance, the intensity differences between the forecasts and the labels of all the 5 models increase. Additionally, each model exhibits unique patterns. Notably, the AM mean values of the differences are

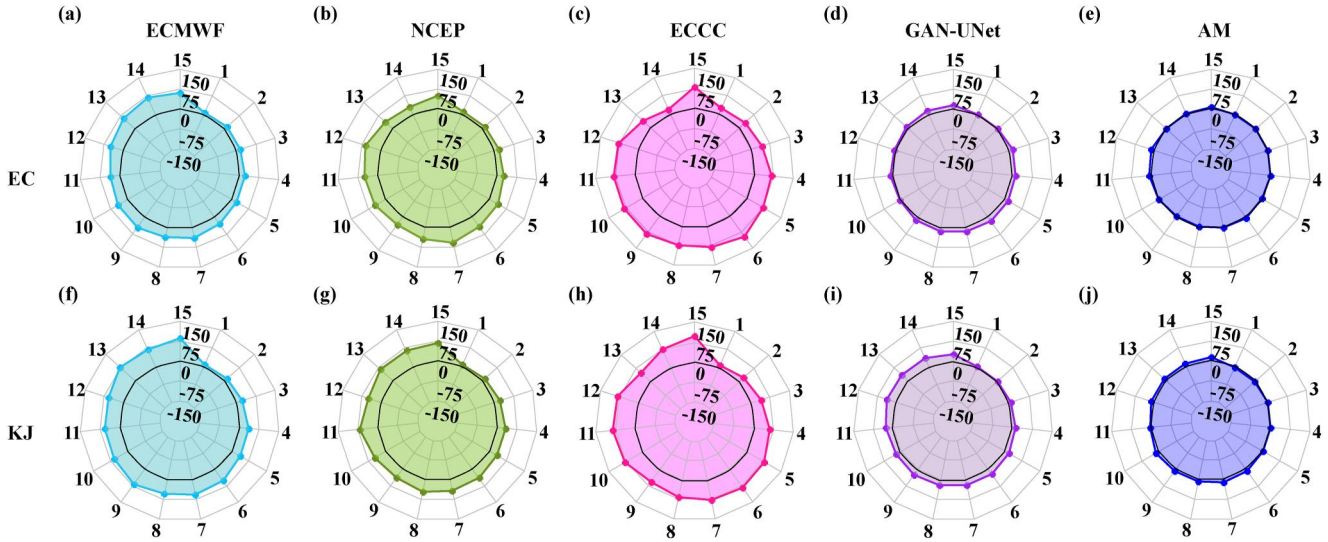

Figure 9. As in Figure 6, but all the forecast and label grids are replaced with 825 kg m^{-1} s^{-1} when IVT > 825 kg m^{-1} s^{-1} at the corresponding grids in the labels.

the least, while those of the GAN-UNet model are slightly greater than those of ECMWF but smaller than those of NCEP and ECCC, with the exception of the final bin.

It is worth mentioning that compared to the mean of the differences, RMSE does have some advantages in evaluating the gap between forecasts and labels. The major distinction lies in that, if only the mean of differences is used to assess the disparity among multiple data points, positive and negative errors may offset each other. However, as illustrated in Figures 7 and 8, employing the mean of differences can demonstrate whether the overall forecast values are biased toward being larger or smaller relative to the labels, which is precisely what we aim to convey. To demonstrate that small differences are not due to the cancellation of positive and negative errors, we have calculated the RMSE, and the corresponding figures are provided in Figures S1 and S2 in Supporting Information S1.

In summary, the forecast results of AR occurrence accuracy, AR shape and intensity show that AM has the best forecast effect on AR events, followed by ECMWF, GAN-UNet, NCEP, and ECCC. One potential reason why NCEP and ECCC perform worse than the ECMWF could be the observation gaps for ARs. NCEP GFS assimilates less data than ECMWF does, causing poor initial conditions. Such deficiency in initial analysis can be projected to poor forecast skills (Geer et al., 2018; Zheng et al., 2021).

3.5. Cases of the Forecast Results

To demonstrate the forecasts more intuitively and confirm the effectiveness of the forecasts, Figures 10 and 11 present two cases of the AR event forecast results generated by different forecast models occurring in EC and KJ, respectively. Selection is subject to the following two criteria: First, remove the uncontroversial AR forecast events, that is, the 5 models mentioned in the study can forecast AR events well. The reason is that we mainly want to show the variability of the models' forecast of controversial events. Second, we do select two cases where GAN-UNet performs better. Through the observation of the forecasted events, we conclude that among controversial forecasted events, GAN-UNet tends to false alarm the AR events that do not occur, while ECMWF tends to miss the events that occur. In general, missing is more harmful than false alarming. In order to highlight the advantages of AM and GAN-UNet methods, we select two occurring events to demonstrate the performance of our proposed method. In the second case, AM and GAN-UNet correctly forecast its occurrence, while both ECMWF and NCEP miss it at the first lead time.

Both the figures indicate that the uncertainty of the forecast and the variability increase as the lead times progress. Specifically, in Figure 10, all forecast results display an IOU above 0.8 when the lead time is 1 day, indicating a relatively accurate forecast of the AR event. The IOUs, ranked from largest to smallest, are AM (0.87), ECMWF (0.86), NCEP (0.844), ECCC (0.838), and GAN-UNet (0.81). The spatial maps also show a good forecast of the

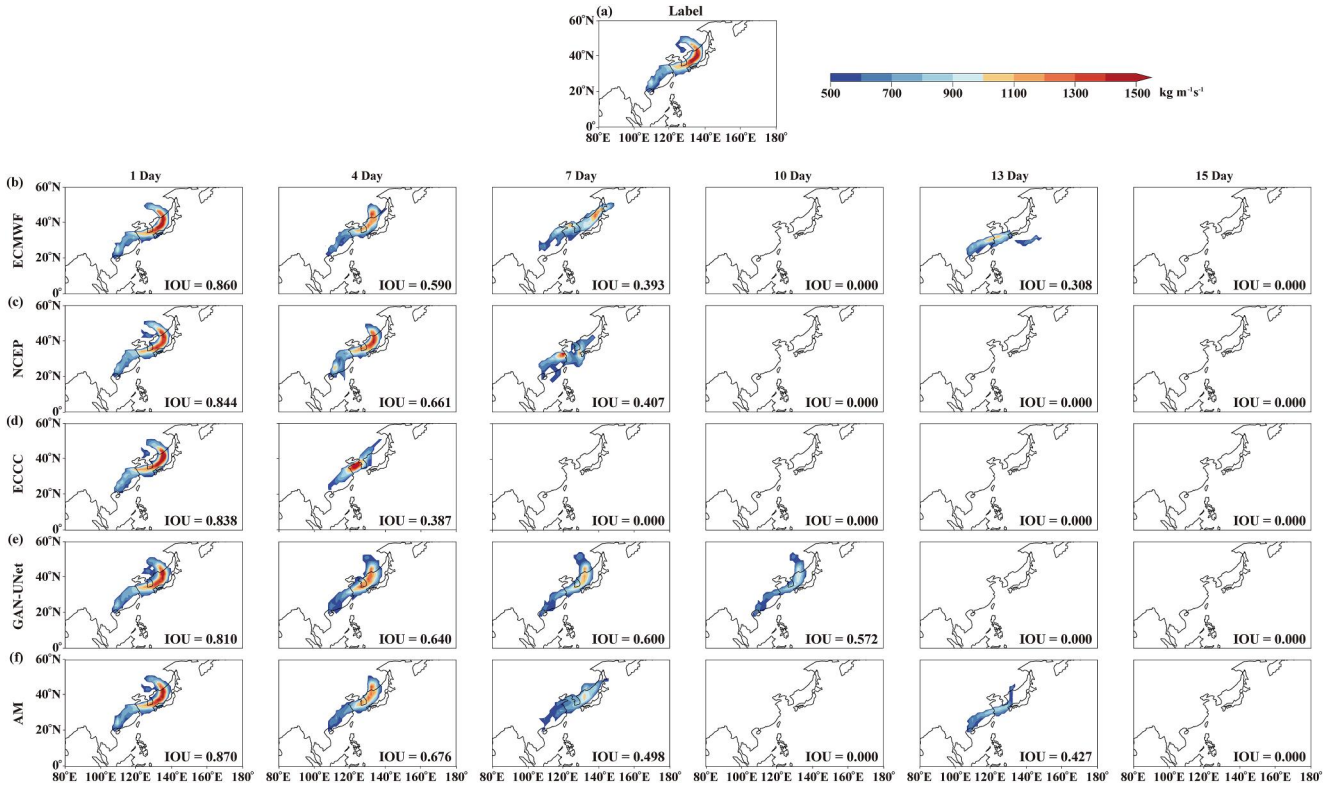

Figure 10. Case study for 02 Jul 2013 1200 UTC with one AR event occurring in EC with (a) the label, and forecasts of (b) ECMWF, (c) NCEP, (d) ECCC, (e) GAN-UNet, and (f) AM. Each column represents the results generated by 1 day, 4 days, 7 days, 10 days, 13 day, and 15 days before the occurrence of this case. For example, the second column data in (b) represents the forecast result obtained at four days before 02 Jul 2013 1200 UTC (i.e., 30 May 2013 1200 UTC) as the initial forecast time.

AR event intensity. However, when the lead times extend to 4 days, all the 5 models' IOUs drop to approximately 0.6, while the performance of AM is still the best. When lead time is 7 days, although all the 5 models forecast the AR event occurrence correctly, the results for all the models except GAN-UNet differ significantly from the label, with the shape and intensity of the AR event shifting. Despite some models forecasting the event beyond 10–15 days, the shapes are markedly disparate with the label and can be deemed as random results. In Figure 11, the AR event occurring in KJ is depicted. When the lead time is 1 day, AM, GAN-UNet, and ECCC correctly forecast the event, with AM having the highest IOU (0.897), followed by GAN-UNet (0.86) and ECCC (0.827). ECMWF and NCEP fail to forecast the AR event. This may be attributed to the definition of the AR event, that is, the AR event needs to cover more than 50% of the key region. Specifically, ECMWF and NCEP may forecast the occurrence of the AR event in East Asia, but since it covers less than 50% areas in KJ, we consider the AR event not to occur there. When the lead time is 4 days, 5 models all forecast the AR event accurately, with AM performing best (0.799), followed by GAN-UNet (0.738), NCEP (0.701), ECCC (0.682), and ECMWF (0.668). Overall, the IOUs obtained by different models fluctuate slightly in a single case, with GAN-UNet and AM consistently demonstrating superior performance, frequently surpassing the state-of-the-art NWP model, especially at the early lead times. However, the reliability of all the 5 models' performance deteriorates when the lead time continue.

From these two cases, it can be seen that the forecast performance of GAN-UNet is comparable to that of NWPs, and the performance of AM is the best, particularly at the early lead times, which is consistent with the results presented earlier in this paper.

4. Discussion and Conclusions

In this paper, we have developed a DL model (GAN-UNet) to forecast the AR events by forecasting the full-field IVT of East Asia with three inputs (IVT, U850, and V850). We make two key improvements to the model architecture: (a) develop a stacked GANs consisting of two generators and a discriminator to better capture the

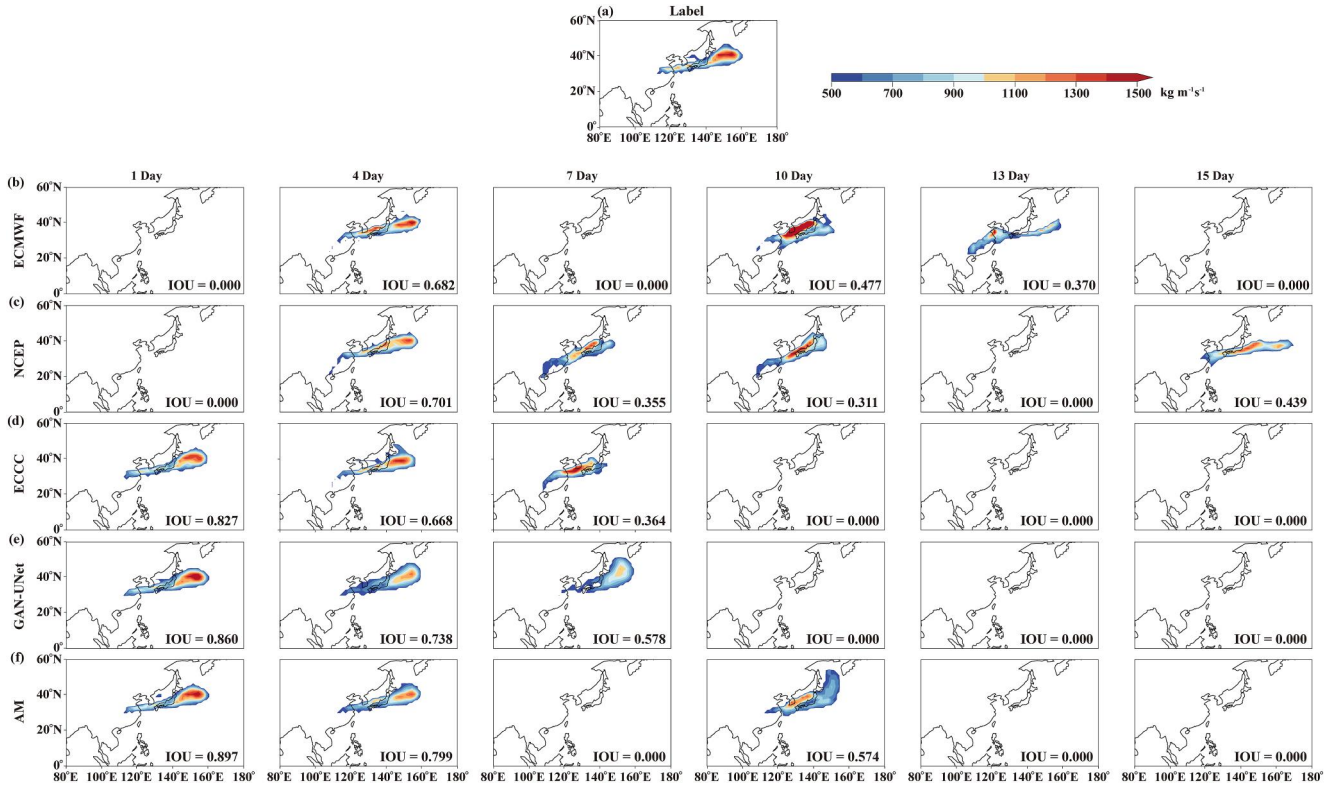

Figure 11. As in Figure 10, but for case study for 27 Jun 2018 1200 UTC with one AR event occurring in KJ.

details of the forecasts, and (b) use the strategy which is called hierarchical temporal aggregation to avoid the error accumulation in the process of multiple forecasts. In addition, in order to better using DL to improve the forecast accuracy of NWP, we have weighted average the results of GAN-UNet and ECMWF models and named it as AM.

GAN-UNet successfully forecast the occurrence, position, and intensity except for the largest IVT bin of the AR events in EC and KJ, with comparable accuracy to the state-of-the-art NWP models achieving in TIGGE. Furthermore, for the forecast of the AR event occurrence and position, AM outperforms all the models used in this study during the early lead times, indicating that GAN-UNet can improve the results of NWP model. Specifically, at the early lead times, in the occurrence and position forecast, AM performance is the best, GAN-UNet outperforms NCEP and ECCC but slightly lags behind ECMWF. At the late lead times, both GAN-UNet and AM are notably better than other NWP models, although the results do not maintain sufficient accuracy. For the AR event intensity forecast, we find that in all the 5 model results, the forecasts for small values are overestimated and the forecasts for large values are underestimated. However, GAN-UNet and AM forecasts show greater consistency and accuracy across all bins, except for the largest bin, where the differences are high for all the 5 models. After replacing the values of the largest bin with a fixed value, GAN-UNet and AM have the least differences in their intensity forecasts compared to the labels.

In general, DL presents several advantages over NWPs. Specifically, (a) the trained DL model can emulate specific modules or processes of NWP models, enhancing accuracy and timeliness. (b) Detection of extreme cases holds paramount importance for disaster prevention and emergency decision-making. Data-driven methods within DL can provide forecasts within minutes of receiving new data, potentially better suiting the requirements of highly responsive forecasting services compared to traditional theory-driven NWPs. (c) Supervised and semi-supervised DL can overcome the limitations of threshold-based conventional detection approaches (Ren et al., 2021; Tian et al., 2023). However, DL is known to have some well-recognized drawbacks, such as poor interpretability and a lack of physical constraints, areas currently receiving focused attention in the DL community. Additionally, addressing corrections/improvements for extreme cases from Machine Learning and DL methods poses challenges, as evident in existing research (Johnson & Khoshgoftaar, 2019; Moniz et al., 2018; Ribeiro & Moniz, 2020). The fundamental challenge in DL-based extreme cases forecasting arises from the rarity

of such cases, a phenomenon known as class imbalance, representing one of the major challenges in machine learning and DL. Highly imbalanced data introduces complexity, as most learners tend to exhibit bias toward the majority class, potentially overlooking the minority class, especially in extreme cases.

This study could be improved in a number of ways. First, while we have identified that AM outperforms ECMWF and GAN-UNet in forecasting the occurrence and position of AR events, we have not quantitatively analyzed the reasons for this phenomenon. Given the varying gaps between GAN-UNet and ECMWF forecasts and the labels, it may be beneficial to consider different weighting strategies. Second, all the 5 models exhibit poor performance in forecasting the maximum values of AR events, which are often associated with severe precipitation and have significant implications for human activities. This may be attributed to the highly imbalanced of the input data and the extensively debated limitation of neural networks: spectral bias (Mojgani et al., 2023). Future studies may consider incorporating algorithms for forecasting extremely rare events (Mojgani et al., 2023; Pickering et al., 2022) to improve the forecast of maximum values in AR events.

Future work is expected to discuss the challenges of accurately forecasting the IVT maximums in AR events, which are common difficulties for most models. We emphasize the importance of accurate forecasting of extreme values for preventing major disasters. Besides, in addition to AR events, we are also concerned about the forecast of AR-related precipitation. In the next work, we hope to get better forecast results of AR-related precipitation and check if they are highly correlated with AR intensity. Additionally, future work will consider post-processing or downscaling of existing forecast data, which would have important implications for disaster prevention and management.

Data Availability Statement

The Pytorch codes for training GAN-UNet is available on (Tian & Zhao, 2023). ERA5 reanalysis data can be downloaded from the European Centre for Medium-Range Weather Forecasts (ECMWF, 2022). The forecast data used in this study to compare our method are the Observing System Research and Predictability Experiment Interactive Grand Global Ensemble (TIGGE, 2008). We thank three anonymous reviewers for their thoughtful comments, which greatly improved the paper.

Acknowledgments
This work was supported by National Key R&D Program of China (2023YFF0805100), Laoshan Laboratory (No.LSKJ202202600), and Shandong Natural Science Foundation Project (ZR2019ZD12).

References

AghaKouchak, A., Chiang, F., Huning, L. S., Love, C. A., Mallakpour, I., Mazdiyasni, O., et al. (2020). Climate extremes and compound hazards in a warming world. *Annual Review of Earth and Planetary Sciences*, *48*(1), 519–548. https://doi.org/10.1146/annurev-earth-071719-055228

Alley, R. B., Emanuel, K. A., & Zhang, F. (2019). Advances in weather prediction. *Science*, *363*(6425), 342–344. https://doi.org/10.1016/j.worlddev.2022.106066

Arabzadeh, A., Ehsani, M. R., Guan, B., Heflin, S., & Behrangi, A. (2020). Global intercomparison of atmospheric rivers precipitation in remote sensing and reanalysis products. *Journal of Geophysical Research: Atmospheres*, *125*(21). https://doi.org/10.1029/2020JD033021

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

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2022). Pangu-weather: A 3D high-resolution model for fast and accurate global weather forecast. arXiv:2211.02556v1.

Chapman, W. E., Delle Monache, L., Alessandrini, S., Subramanian, A. C., Ralph, F. M., Xie, S. P., et al. (2022). Probabilistic predictionss from deterministic atmospheric river forecasts with deep learning. *Monthly Weather Review*, *150*(1), 215–234. https://doi.org/10.1175/MWR-D-21-0106.1

Chapman, W. E., Subramanian, A. C., Delle Monache, L., Xie, S. P., & Ralph, F. M. (2019). Improving atmospheric river forecasts with machine learning. *Geophysical Research Letters*, *46*(17–18), 10627–10635. https://doi.org/10.1029/2019GL083662

Chen, G., & Wang, W. C. (2022). Short-term precipitation prediction for contiguous United States using deep learning. *Geophysical Research Letters*, *49*(8). https://doi.org/10.1029/2022GL097904

Chen, X., Leung, L. R., Gao, Y., Liu, Y., Wigmosta, M., & Richmond, M. (2018). Predictability of extreme precipitation in western US watersheds based on atmospheric river occurrence, intensity, and duration. *Geophysical Research Letters*, *45*(21), 11–693. https://doi.org/10.1029/2022GL097904

DeFlorio, M. J., Waliser, D. E., Guan, B., Lavers, D. A., Ralph, F. M., & Vitart, F. (2018). Global assessment of atmospheric river prediction skill. *Journal of Hydrometeorology*, *19*(2), 409–426. https://doi.org/10.1175/JHM-D-17-0135.1

Durugkar, I., Gemp, I., & Mahadevan, S. (2016). Generative multi-adversarial networks. arXiv preprint arXiv:1611.01673.

ECMWF. (2022). ERA5 hourly data on pressure levels from 1959 to present. [Dataset]. Retrieved from https://cds.climate.copernicus.eu/

Espinoza, V., Waliser, D. E., Guan, B., Lavers, D. A., & Ralph, F. M. (2018). Global analysis of climate change projection effects on atmospheric rivers. *Geophysical Research Letters*, *45*(9), 4299–4308. https://doi.org/10.1029/2017GL076968

Geer, A. J. (2021). Learning earth system models from observations: Machine learning or data assimilation? *Philosophical Transactions of the Royal Society A*, *379*(2194), 20200089. https://doi.org/10.1098/rsta.2020.0089

Geer, A. J., Lonitz, K., Weston, P., Kazumori, M., Okamoto, K., Zhu, Y., et al. (2018). All-sky satellite data assimilation at operational weather forecasting centres. *Quarterly Journal of the Royal Meteorological Society*, *144*(713), 1191–1217. https://doi.org/10.1002/qj.3202

Ghosh, A., Kulharia, V., Namboodiri, V. P., Torr, P. H., & Dokania, P. K. (2018). Multi-agent diverse generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 8513–8521).

Gimeno, L., Nieto, R., Vázquez, M., & Lavers, D. A. (2014). Atmospheric rivers: A mini-review. *Frontiers in Earth Science*, *2*, 2. https://doi.org/10.3389/feart.2014.00002

Gong, B., Langguth, M., Ji, Y., Mozaffari, A., Stadtler, S., Mache, K., & Schultz, M. G. (2022). Temperature forecasting by deep learning methods. *Geoscientific Model Development*, *15*(23), 8931–8956. https://doi.org/10.5194/gmd-15-8931-2022

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139–144. https://doi.org/10.1145/3422622

Guan, B., & Waliser, D. E. (2015). Detection of atmospheric rivers: Evaluation and application of an algorithm for global studies. *Journal of Geophysical Research: Atmospheres*, *120*(24), 12514–12535. https://doi.org/10.1002/2015JD024257

Guan, B., Waliser, D. E., & Ralph, F. M. (2023). Global application of the atmospheric river scale. *Journal of Geophysical Research: Atmospheres*, *128*(3), e2022JD037180. https://doi.org/10.1029/2022JD037180

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Higgins, T. B., Subramanian, A. C., Graubner, A., Kapp-Schwoerer, L., Watson, P. A., Sparrow, S., et al. (2023). Using deep learning for an analysis of atmospheric rivers in a high-resolution large ensemble climate data set. *Journal of Advances in Modeling Earth Systems*, *15*(4), e2022MS003495. https://doi.org/10.1029/2022MS003495

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*(1), 1–54. https://doi.org/10.1186/s40537-019-0192-5

Kamae, Y., Imada, Y., Kawase, H., & Mei, W. (2021). Atmospheric rivers bring more frequent and intense extreme rainfall events over East Asia under global warming. *Geophysical Research Letters*, *48*(24), e2021GL096030. https://doi.org/10.1029/2021GL096030

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). Progressive growing of glans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196.

Kashinath, K., Kashinath, K., Mudigonda, M., Kim, S., Kapp-Schwoerer, L., Graubner, A., et al. (2021). ClimateNet: An expert-labeled open dataset and deep learning architecture for enabling high-precision analyses of extreme weather. *Geoscientific Model Development*, *14*(1), 107–124. https://doi.org/10.5194/gmd-14-107-2021

Kim, J., Moon, H., Guan, B., Waliser, D. E., Choi, J., Gu, T. Y., & Byun, Y. H. (2021). Precipitation characteristics related to atmospheric rivers in East Asia. *International Journal of Climatology*, *41*(S1), E2244–E2257. https://doi.org/10.1002/joc.6843

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Pritzel, A., et al. (2022). GraphCast: Learning skillful medium-range global weather forecasting. arXiv preprint arXiv:2212.12794.

Lavers, D. A., Pappenberger, F., & Zsoter, E. (2014). Extending medium-range predictability of extreme hydrological events in Europe. *Nature Communications*, *5*(1), 1–7. https://doi.org/10.1038/ncomms6382

Lavers, D. A., Ralph, F. M., Richardson, D. S., & Pappenberger, F. (2020). Improved forecasts of atmospheric rivers through systematic reconnaissance, better modelling, and insights on conversion of rain to flooding. *Communications Earth & Environment*, *1*(1), 1–7. https://doi.org/10.1038/s43247-020-00042-1

Lavers, D. A., & Villarini, G. (2013). The nexus between atmospheric rivers and extreme precipitation across Europe. *Geophysical Research Letters*, *40*(12), 3259–3264. https://doi.org/10.1002/grl.50636

Li, J., Zhao, Y., Chen, D., Kang, Y., & Wang, H. (2021). Future precipitation changes in three key sub-regions of East Asia: The roles of thermodynamics and dynamics. *Climate Dynamics*, *59*(5–6), 1–22. https://doi.org/10.1007/s00382-021-06043-w

Li, J., Zhao, Y., & Tang, Z. (2020). Projection of future summer precipitation over the Yellow River Basin: A moisture budget perspective. *Atmosphere*, *11*(12), 1307. https://doi.org/10.3390/atmos11121307

Liang, J., & Yong, Y. (2021). Climatology of atmospheric rivers in the Asian monsoon region. *International Journal of Climatology*, *41*(S1), E801–E818. https://doi.org/10.1002/joc.6729

Liang, J., Yong, Y., & Hawcroft, M. K. (2022). Long-term trends in atmospheric rivers over East Asia. *Climate Dynamics*, *1*(3–4), 643–724. https://doi.org/10.1007/s00382-022-06339-5

Mahoney, K., Jackson, D. L., Neiman, P., Hughes, M., Darby, L., Wick, G., et al. (2016). Understanding the role of atmospheric rivers in heavy precipitation in the southeast United States. *Monthly Weather Review*, *144*(4), 1617–1632. https://doi.org/10.1175/MWR-D-15-0279.1

Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., et al. (2021). Climate change 2021: The physical science basis. *Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change*. Retrieved from https://www.ipcc.ch/report/ar6/wg1/#SPM

Mojgani, R., Chattopadhyay, A., & Hassanzadeh, P. (2023). Interpretable structural model error discovery from sparse assimilation increments using spectral bias-reduced neural networks: A quasi-geostrophic turbulence test case. arXiv preprint arXiv:2309.13211.

Molteni, F., Buizza, R., Palmer, T. N., & Petroliagis, T. (1996). The ECMWF ensemble prediction system: Methodology and validation. *The Quarterly Journal of the Royal Meteorological Society*, *122*(529), 73–119. https://doi.org/10.1002/qj.49712252905

Moniz, N., Ribeiro, R., Cerqueira, V., & Chawla, N. (2018). *Smoteboost for regression: Improving the prediction of extreme values*. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 150–159). IEEE. https://doi.org/10.1109/DSAA.2018.00025

Mundhenk, B. D., Barnes, E. A., & Maloney, E. D. (2016). All-season climatology and variability of atmospheric river frequencies over the North Pacific. *Journal of Climate*, *29*(13), 4885–4903. https://doi.org/10.1175/JCLI-D-15-0655.1

Nardi, K. M., Barnes, E. A., & Ralph, F. M. (2018). Assessment of numerical weather prediction model reforecasts of the occurrence, intensity, and location of atmospheric rivers along the West Coast of North America. *Monthly Weather Review*, *146*(10), 3343–3362. https://doi.org/10.1175/MWR-D-18-0060.1

Nayak, M. A., Villarini, G., & Lavers, D. A. (2014). On the skill of numerical weather prediction models to forecast atmospheric rivers over the central United States. *Geophysical Research Letters*, *41*(12), 4354–4362. https://doi.org/10.1002/2014GL060299

Pan, M., & Lu, M. (2019). A novel atmospheric river identification algorithm. *Water Resources Research*, *55*(7), 6069–6087. https://doi.org/10.1029/2018WR024407

Park, Y. Y., Buizza, R., & Leutbecher, M. (2008). TIGGE: Preliminary results on comparing and combining ensembles. *Quarterly Journal of the Royal Meteorological Society: A Journal of the Atmospheric Sciences, Applied Meteorology and Physical Oceanography*, *134*(637), 2029–2050. https://doi.org/10.1002/qj.334

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., et al. (2022). Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. arXiv preprint arXiv:2202.11214.

Payne, A. E., Demory, M. E., Leung, L. R., Ramos, A. M., Shields, C. A., Rutz, J. J., et al. (2020). Responses and impacts of atmospheric rivers to climate change. *Nature Reviews Earth and Environment*, *1*(3), 143–157. https://doi.org/10.1038/s43017-020-0030-5

Pickering, E., Guth, S., Karniadakis, G. E., & Sapsis, T. P. (2022). Discovering and forecasting extreme events via active learning in neural operators. *Nature Computational Science*, *2*(12), 823–833. https://doi.org/10.1038/s43588-022-00376-0

Ralph, F. M., Neiman, P. J., Wick, G. A., Gutman, S. I., Dettinger, M. D., Cayan, D. R., & White, A. B. (2006). Flooding on California's Russian River: Role of atmospheric rivers. *Geophysical Research Letters*, *33*(13), L13801. https://doi.org/10.1029/2006GL026689

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

Reid, K. J., King, A. D., Lane, T. P., & Short, E. (2020). The sensitivity of atmospheric river identification to integrated water vapor transport threshold, resolution, and regridding method. *Journal of Geophysical Research: Atmospheres*, *125*(20), 1–15. https://doi.org/10.1029/2020JD032897

Ren, X., Li, X., Ren, K., Song, J., Xu, Z., Deng, K., & Wang, X. (2021). Deep learning-based weather prediction: A survey. *Big Data Research*, *23*, 100178. https://doi.org/10.1016/j.bdr.2020.100178

Ribeiro, R. P., & Moniz, N. (2020). Imbalanced regression and extreme value prediction. *Machine Learning*, *109*(9–10), 1803–1835. https://doi.org/10.1007/s10994-020-05900-9

Ronneberger, O., Fischer, P., & Brox, T. (2015). *U-net: Convolutional networks for biomedical image segmentation*. In *Paper Presented at 18th International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28

Scoccimarro, E., Gualdi, S., Bellucci, A., Zampieri, M., & Navarra, A. (2013). Heavy precipitation events in a warmer climate: Results from CMIP5 models. *Journal of Climate*, *26*(20), 7902–7911. https://doi.org/10.1175/JCLI-D-12-00850.1

Slinskey, E. A., Loikith, P. C., Waliser, D. E., Guan, B., & Martin, A. (2020). A climatology of atmospheric rivers and associated precipitation for the seven US national climate assessment regions. *Journal of Hydrometeorology*, *21*(11), 2439–2456. https://doi.org/10.1175/JHM-D-20-0039.1

Swinbank, R., Kyouda, M., Buchanan, P., Froude, L., Hamill, T. M., Hewson, T. D., et al. (2016). The TIGGE project and its achievements. *Bulletin of the American Meteorological Society*, *97*(1), 49–67. https://doi.org/10.1175/BAMS-D-13-00191.1

Tian, Y., & Zhao, Y. (2023). AR forecast method. [ComputationalNotebook]. Science Data Bank. https://www.scidb.cn/s/yEnmum

Tian, Y., Zhao, Y., Son, S. W., Luo, J. J., Oh, S. G., & Wang, Y. (2023). A deep-learning ensemble method to detect atmospheric rivers and its application to projected changes in precipitation regime. *Journal of Geophysical Research: Atmospheres*, *128*(12), e2022JD037041. https://doi.org/10.1029/2022JD037041

TIGGE. (2008). The forecast data on pressure levels. [Dataset]. Retrieved from https://apps.ecmwf.int/datasets/data/tigge

Waliser, D., & Guan, B. (2017). Extreme winds and precipitation during landfall of atmospheric rivers. *Nature Geoscience*, *10*(3), 179–183. https://doi.org/10.1038/ngeo2894

Wang, T., Wei, K., & Ma, J. (2021). Atmospheric rivers and mei-yu rainfall in China: A case study of summer 2020. *Advances in Atmospheric Sciences*, *38*(12), 2137–2152. https://doi.org/10.1007/s00376-021-1096-9

Wick, G. A., Neiman, P. J., Ralph, F. M., & Hamill, T. M. (2013). Evaluation of forecasts of the water vapor signature of atmospheric rivers in operational numerical weather prediction models. *Weather and Forecasting*, *28*(6), 1337–1352. https://doi.org/10.1175/WAF-D-13-00025.1

Zhang, P., Chen, G., Ting, M., Ruby Leung, L., Guan, B., & Li, L. (2023). More frequent atmospheric rivers slow the seasonal recovery of Arctic sea ice. *Nature Climate Change*, *13*(3), 266–273. https://doi.org/10.1038/s41558-023-01599-3

Zhao, Y., Park, C., & Son, S.-W. (2023). Importance of diabatic heating for the eastward-moving heavy rainfall events along the Yangtze River, China. *Journal of the Atmospheric Sciences*, *80*(1), 151–165. https://doi.org/10.1175/JAS-D-21-0321.1

Zheng, M., Delle Monache, L., Cornuelle, B. D., Ralph, F. M., Tallapragada, V. S., Subramanian, A., et al. (2021). Improved forecast skill through the assimilation of dropsonde observations from the Atmospheric River Reconnaissance program. *Journal of Geophysical Research: Atmospheres*, *126*(21), e2021JD034967. https://doi.org/10.1029/2021JD034967

Zhu, Y., & Newell, R. E. (1994). Atmospheric rivers and bombs. *Geophysical Research Letters*, *21*(18), 1999–2002. https://doi.org/10.1029/94GL01710