CLGAN: A GAN-based video prediction model for precipitation nowcasting

Yan Ji[1,2], Bing Gong[2], Michael Langguth[2], Amirpasha Mozaffari[2], and Xiefei Zhi[1]

[1]Nanjing University of Information Science and Technology, 210044 Nanjing, China
[2]Jülich Supercomputing Centre, Forschungszentrum Jülich, 52425 Jülich, Germany

Correspondence: b.gong@fz-juelich.de

Abstract. The prediction of precipitation patterns up to two hours ahead, also known as precipitation nowcasting, at high spatio-temporal resolutions is of great relevance in weather-dependent decision making and early warning systems. In this study, we are aiming to provide an efficient and easy-to-understand deep neural network - CLGAN, to improve the nowcasting skills of heavy precipitation events. The model constitutes a Generative Adversarial Network (GAN) architecture, whose generator is built upon a u-shaped encoder-decoder network (U-Net) and equips with recurrent LSTM cells to capture spatio-temporal features. The optical flow model DenseRotation and the competitive video prediction models ConvLSTM and PredRNN-V2 are used as the competitors. A series of evaluation metrics, including the root mean square error, the critical success index, the fractions skill score and object-based diagnostic evaluation, are utilized for a comprehensive comparison against competing baseline models. We show that CLGAN outperforms the competitors in terms of scores for dichotomous events and object-based diagnostics. A sensitivity analysis on the weight of the GAN-component indicates that the GAN-based architecture helps to capture heavy precipitation events. The results encourage future work based on the proposed CLGAN architecture to improve the precipitation nowcasting and early-warning systems.

1 Introduction

Heavy precipitation can lead to numerous hazards, cause damages to infrastructure, and even increase risk of human life (Ganguly and Bras, 2003; Vasiloff et al., 2007; Li et al., 2021). Accurate short-term predictions of precipitation events at high spatio-temporal resolutions, also known as precipitation nowcasting, are therefore critical in establishing early-warning systems. These warning systems can in turn promote authorities in weather-dependent decision-making and enhance risk-governance capabilities (Dixon and Wiener, 1993; Johnson et al., 1998; Bowler et al., 2006).

Current precipitation nowcasting systems mainly rely on convective-permitting numeric weather prediction (NWP) or on extrapolation techniques of precipitation patterns with the help of composite radar observations. However, NWP models suffer from difficulties in capturing these patterns in the nowcasting time range due to the spin-up effect and the challenges of handling non-Gaussian data in assimilation (Ravuri et al., 2021). Also, a quick model run cycle would be required. For instance, ICON-D2 only initializes every 3 hours (Matsunobu et al., 2022), which makes it impossible to get quick updates in light of rapidly growing precipitation patterns. Observation-based extrapolation methods, such as optical flow, are commonly superior to NWP

25 models for precipitation nowcasting, but also fail to capture the underlying non-linear processes of precipitation formation, e.g. secondary triggering and aggregation (Xie et al., 2019).

Deep neural networks have gained increasing attention in the meteorological community over the last few years (Reichstein et al., 2019; Schultz et al., 2021). The growing interest can be attributed to the success stories in other domains where deep learning (DL) proves to leverage high-level information from complex and highly non-linear data in several applications,

30 such as autonomous driving (Hu et al., 2020), anomaly detection (Liu et al., 2018), and semantic segmentation (Garcia-Garcia et al., 2018). Recently, video prediction models, developed in the computer vision community, have been explored for precipitation nowcasting. Contemporary studies mainly applied model architectures based on u-shaped convolutional networks (U-net; e.g. Ayzel et al., 2020; Ronneberger et al., 2015), convolutional long short-term memory cells (ConvLSTM; e.g. Shi et al., 2015), generative (see e.g. Ravuri et al., 2021) as well as attention models (see e.g. Sønderby et al., 2020). U-Nets are

35 thereby considered to be beneficial since they are capable to extract multi-scale features of the atmospheric processes (Ronneberger et al., 2015). To also explicitly capture temporal dependencies in the underlying formation process of precipitation, recurrent ConvLSTM models are an appealing choice (Shi et al., 2015). Thus, combining convolutional and recurrent networks with ConvLSTM-layers is advantageous in generating stable precipitation nowcasting by encoding the spatial and temporal dependency from the historical frames.

40 Nevertheless, these models have problems with handling the statistical nature of precipitation, especially when a pixel-wise loss function is applied for the optimization process during training (Shi et al., 2017; Ayzel et al., 2020). Although log transformation, importance sampling and weighting towards heavier precipitation targets are appropriate to govern the right-skewed Gamma distribution of precipitation rates (e.g. Ravuri et al., 2021), the inherent uncertainty in quasi-chaotic processes at meso-scale typically leads to unrealistically smooth precipitation patterns in the forecasts. While this issue is well known

45 in many other video prediction tasks (Mathieu et al., 2015; Ebert et al., 2017), it is of particular relevance in precipitation nowcasting. The high spatio-temporal variability seen in the observational (real) data cannot be maintained and thus, heavy precipitation events are barely captured by models applying a pixel-wise loss. Generative models, which train a generator and a discriminator adversarially (GAN models) are considered to be a potential solution for such applications (Goodfellow et al., 2020). By forcing the generator to fool the discriminator, which aims to distinguish between real and generated data, these

50 models succeed in maintaining the statistical properties of the underlying data (Oprea et al., 2020).

Although great progress has been achieved in a series of recent studies (e.g. Ravuri et al., 2021; Gong et al., 2022), there is controversy regarding how different components of sophisticated model architectures contribute to the predictions. Motivated by this, we build a simple, but efficient and easy-to-understand video prediction model, CLGAN (Convolutional Long short-term memory Generative Adversarial Network; see Fig. 1), for the nowcasting task. CLGAN is proposed to leverage the

55 advantages of different DL model architectures. The generator combines the U-Net with a ConvLSTM cell to abstract spatial features on multiple scales, while the temporal dependency of precipitation patterns is also preserved. The generator network is then trained adversarially to attain precipitation forecasts resembling observed data. For our nowcasting application, we deploy a gridded dataset with a temporal resolution of 10 minutes aggregated from automatic weather station (AWS) gauges over Guizhou, China. The predictive performance of the proposed model architecture is then accessed in a comprehensive evalu-

ation based on metrics designed for precipitation nowcasting. The evaluation also involves a comparison against a simplistic persistence forecast, the conventional optical flow model DenseRotation (Ayzel et al., 2019) as well as two baseline video prediction models, a standard ConvLSTM (Shi et al., 2015) network and an up-to-date competing model PredRNN-v2 (Wang et al., 2021).

With this, the main contributions of our study are:

- An efficient and easy-to-understand architecture CLGAN leveraging the merits of U-Net, ConvLSTM and GAN models is proposed to generate perceptually realistic precipitation forecasts.

- A new 10-minute-level precipitation dataset based on AWS gauges (Guizhou AWS_ML precipitation dataset) is built for machine learning experiments.

- Nowcasting of heavy precipitation events is improved with comprehensive verification.

- A sensitivity analysis is performed to access the importance of adversarial training for generating forecasts with closer statistical properties of the observed precipitation.

2 Related work and baseline models

Conventional Methods. The simplest approach to generate precipitation 'forecasts' is to deploy the Eulerian persistence. For this, the most recent available observation, usually a radar composite, is used and then replicated several times for the future steps. This approach is quite accurate for very short lead times, but obviously fails to provide meaningful forecasts in a quickly evolving system for time scales beyond several minutes such as the atmosphere. Thus, the related forecast quality can be considered as the minimum level for a prediction model to be useful.

Conventional precipitation nowcasting systems typically use a Lagrangian framework to predict the development of precipitation patterns. Although this framework often assumes persistence of the precipitation features' intensity and displacement, it is still capable to outperform mesoscale NWP models in precipitation nowcasting (Sun et al., 2014). The Lagrangian method applies a two-step approach where the precipitation features are first tracked and then extrapolated to future time steps (Austin and Bellon, 1974). Typically, the tracking step is accomplished with the help of optical flow methods that infer the motion of patterns from consecutive images. For precipitation nowcasting, radar composite images are subject to a tracking algorithm such as cross-correlation tracking (Rinehart and Garvey, 1978; Grecu and Krajewski, 2000; Zahraei et al., 2012) or centroid tracking techniques (Zahraei et al., 2013). The tracked objects are then applied to different extrapolation schemes, e.g. image warping (Wolberg, 1990), constant-vector advection (Bowler et al., 2004) or semi-Lagrangian schemes (Germann and Zawadzki, 2002). With this two-step approach, several operational precipitation nowcasting systems have been established over the globe in the last three decades, such as the Thunderstorm Identification Tracking Nowcasting (TITAN; Dixon and Wiener, 1993), the Storm Cell Identification and Tracking (SCIT; Johnson et al., 1998), and the Short-Term Ensemble Prediction System (STEPS; Bowler et al., 2006) (see Wilson et al., 2010, for a review on operational systems).

Recently, Ayzel et al. (2019) implemented a set of advanced optical flow models into an open-source Python library called *rainymotion*. Two different groups of methods are part of this library from which we select in the DenseRotation model that performs best in their study. The tracking algorithm of this model is based on the Dense Inverse Search algorithm proposed in Kroeger et al. (2016) providing an estimate of the motion of each pixel based on two consecutive radar images. The extrapolation is then performed with a semi-Lagrangian advection scheme (Germann and Zawadzki, 2002) capable to represent rotational motions.

In our study, the Eulerian persistence model and the DenseRotation model are used to show how well the traditional methods can perform for the precipitation nowcasting task and how much benefit can be further obtained by using DL-based video prediction methods.

Video Prediction Method. As already mentioned, the application of deep learning techniques in the meteorological community has gained momentum over the recent years. In particular, several studies have started to explore these techniques to tackle the precipitation nowcasting problem. By formulating precipitation nowcasting as a sequence prediction task, Shi et al. (2015) proposed a network of ConvLSTM cells which apply a convolution in the recurrent layers of the vanilla LSTM to capture spatio-temporal features in the underlying data. Their 2-layer ConvLSTM network was able to outperform outperform the Variational method for Echoes of Radar by Woo and Wong (ROVER; 2017), an operational precipitation nowcasting system based on optical flow methods with a semi-Lagrangian advection scheme. Shi et al. (2017) further extended the recurrent cells of GRUs with non-local neural connections and proposed the Trajectory GRU (TrajGRU) model which enables learning of location-variant structures of precipitation.

Besides, Wang et al. (2017) advanced the application of ConvLSTM networks and proposed the predictive recurrent neural network (PredRNN). They deployed a stack of recurrent layers that feature a 'zigzag memory flow' and involve an explicit spatio-temporal memory state. In this way, they enable an explicit communication of abstracted spatio-temporal features between different levels of the recurrent network which yields improved precipitation predictions. While this approach already provided promising results, the PredRNN model was updated to PredRNN-V2 (Wang et al., 2021). The updates comprise the implementation of a 'decoupling loss', named ST-LSTM, to enhance featuring of the spatio-temporal variations and a new, improved long-term modeling strategy. The model attains remarkable improvements when applied to multiple datasets including radar observations.

Meanwhile, other network architectures were explored in scope of precipitation nowcasting. One of them is the fully convolutional U-Net architecture, which is a U-shaped hierarchical encoder-decoder network with skip connections. The architecture enables abstraction of features on different spatial scales. Notably, the RainNet-architecture proposed by Ayzel et al. (2020) proved to significantly outperform optical flow based nowcasting methods for weak precipitation events. However, their network tends to provide too smooth precipitation fields and therefore fails to provide added value for more intense precipitation events with a rain rate above $10\,\mathrm{mm/h}$. Recently, a deep generative model for the probabilistic precipitation nowcasting was proposed and showed state-of-the-art performance for the task (Ravuri et al., 2021).

All these studies demonstrate that deep neural networks have the potential to provide added value for precipitation nowcasting. In our study, we focus on further improving the predictions of strong precipitation events and therefore choose a simple

ConvLSTM (Shi et al., 2015) and the advanced PredRNN-v2 (Wang et al., 2021) model for competing with our newly proposed model architecture.

3 Method and Data

3.1 Our model CLGAN

In the following, we present our proposed CLGAN architecture in more detail. Since CLGAN aims to benefit from ConvLSTM-models, the U-Net architecture and GAN models, we first introduce its components separately to provide a deeper understanding and reasoning for the chosen components.

3.1.1 ConvLSTM

The ConvLSTM network was proposed as an extension of LSTM layers which embedded the convolution operation to explicitly encode complex spatio-temporal features in a data sequence. The basic formulas of the ConvLSTM cell which describe the gated update procedure for the hidden and the cell state are provided in Shi et al. (2015) and are therefore not repeated here. The objective function of a ConvLSTM model typically constitutes the classical $\mathcal{L}^2$ reconstruction loss. This loss measures the distance between the predicted and the target (ground truth) data on grid point (or pixel-wise) level and can be written as

$$\mathcal{L}^2(G) = \left\| \overrightarrow{Y}_{t_0+1:t} - \hat{\overrightarrow{X}}_{t_0+1:t} \right\|_2 \tag{1}$$

where $\overrightarrow{Y}$ and $\hat{\overrightarrow{X}}$ are two-dimensional tensors for the ground truth and the predicted data, respectively. t_0 represents the end of the input sequence and t is the forecast time step, so that the model is optimized on the loss over the prediction sequence from $t_0 + 1$ to t. These tensors comprise $w \times h$ grid points in zonal and meridional direction of the domain of interest.

3.1.2 U-Net

The U-Net model was originally applied for biomedical image segmentation (Ronneberger et al., 2015) and is therefore designed as a powerful feature extractor on various spatial scales. As illustrated in Fig. 1a, it can be decomposed into a compressing and an expansive path that are bridged by skip connections. The contracting path can be seen as an encoder which converts the high-resolved data into coarse-grained features using convolutional and pooling layers. The expansive path, acting as a decoder, applies deconvolutional layers to convert back to the original spatial resolution, of which the number of data points are $w \times h$. Usually, several pooling and deconvolutional layers are applied to allow feature extraction on different spatial scales. To avoid the so-called vanishing gradients issue and to allow a direct information flow of specific spatial features, skip connections are implemented at every scale-specific feature extraction level (Drozdzal et al., 2016).

In a video prediction application, the data at time step t enters the encoder to produce a forecast at time step $t + 1$ with the decoder. By doing so, no long-term information is explicitly conveyed as with the ConvLSTM-model. Since heavy precipitation events are rare, but of high relevance for nowcasting, different techniques are usually applied to encourage deep neural networks

in predicting events on the right tail of the underlying PDF. Log-transformation converts the right-skewed Gamma distribution of precipitation data (e.g. RainNet in Ayzel et al., 2020) into a Gaussian-like distribution which puts strong precipitation events closer to the center of mass in probability space. Stronger weighting on higher precipitation rates and importance sampling can further support the optimization efficiency with respect to heavy precipitation events (Ravuri et al., 2021). Nonetheless, U-Nets and ConvLSTM modes still tend to produce too smooth precipitation patterns, thereby failing to capture the relevant strong precipitation events.

3.1.3 Generative adversarial networks

To enforce a closer agreement of the generated data with the ground truth, GAN models were proposed by Goodfellow et al. (2020). A GAN model consists of a generative network G (generator) and a discriminative network D (discriminator) which aims to assign a probability of 1 to real and a probability of 0 to generated data. While the discriminator is optimized to distinguish between both kinds of inputted data, the generator is encouraged to fool the discriminator. Thus, the GAN applies the binary cross-entropy loss as the objective function which enters a minimax-game:

$$G^\star = \arg\min_{G}\max_{D}\mathcal{L}^{GAN}(G,D)$$

$$\text{with } \mathcal{L}^{GAN}(G,D) = \mathbb{E}_{\vec{X}_{1:t}}\left[\log D(\vec{X}_{t_0+1:t})\right] + \mathbb{E}_{\vec{X}_{1:t}}\left[\log(1 - D(G(\vec{X}_{1:t_0})))\right]. \tag{2}$$

Here, the generator is conditioned on the input data sequence $\vec{X}_{1:t_0}$. Since generator and discriminator are trained adversarially, the generator is encouraged to create predictions that share the same statistical properties as the ground truth data. This is considered to be useful for generating realistic precipitation forecasts which should exhibit the high spatial variability of the observed data (Ravuri et al., 2021; Price and Rasp, 2022; Harris et al., 2022).

3.1.4 Convolutional LSTM GAN (CLGAN)

To combine the merits of a GAN model with the strong spatio-temporal feature extraction capacities of U-Nets and ConvLSTM models, we set up the generator G as follows (see Fig. 1a): The generator constitutes a three-level U-Net following Sha et al. (2020). Each level of the encoder comprises two convolutional layers followed by max-pooling with a 2x2-kernel to reduce the spatial dimensionality in the next layer. The number of channels is thereby increased by a factor of two in each level. A ConvLSTM cell with 64 filters is deployed to implement recurrency at the bridge between the encoder and decoder. The decoder then reverts the encoded data to the input resolution with the help of deconvolutional layers. Furthermore, skip connections among the encoder and decoder are added at each level of the U-Net. The discriminator D consists of 3D fully convolutional layers with batch normalization which allow us to encode both, the temporal and spatial dimension of the data sequence. Again, max-pooling is used to compress the data which finally gets concatenated to fully connected layers (see Fig. 1b). The forecast sequence of G, $\hat{\vec{X}}_{t_0+1:t}$, as well as the corresponding ground-truth sequence, $\vec{Y}_{t_0+1:t}$, are taken as the inputs for the discriminator D (see Fig. 1c).

In this study, the generator is trained by combining the adversarial loss $\mathcal{L}^{GAN}$ with the reconstruction $\mathcal{L}^2$-loss:

$$G^\star = (1-\lambda)\mathcal{L}^{GAN}(G,D) + \lambda\mathcal{L}^2(G) \ \text{ with } \lambda \in [0,1]. \tag{3}$$

This ensures that the prediction remains close to the ground truth. The relative weight of the reconstruction loss λ is set to 0.99 which proves to balance the contributions from both loss components in the following experiments. Training of the model is performed with the Adam optimizer (Kingma and Ba, 2014) over eight epochs with a batch size of 32.

3.2 Guizhou AWS_ML precipitation dataset

In addition to the widely used remote sensing data, e.g. radar composite images, measurements from densely-distributed automatic weather stations can serve as an alternative in the data-driven weather forecasting task. In this study, minute-level precipitation measurements by rain gauges of AWS over Guizhou, China (Guizhou AWS_ML precipitation dataset) is collected for the precipitation nowcasting task. Guizhou is a mountainous and rainy province located in southwest China (see Fig. 2a) where mudslides happen frequently during summer-times. For instance, the region was affected by a severe rainstorm in September 2020, when some regions experienced more than $1500\,\text{mm}$ rainfall within 20 days. Accurate precipitation nowcasting, especially for heavy precipitation, is crucial to reduce damages from these events. Hence, the Guizhou AWS_ML precipitation dataset is established for better simulation of precipitation with data-driven approaches. The AWS locations comprise 93 national basic stations and 1740 automatic weather stations (see Fig. 2b). Among other meteorological quantities (2-m temperature, 10-m wind, surface pressure, and relative humidity), the AWSs measure precipitation at a high observation frequency (every minute) and the data is provided between 1st Jan 2015 and 31st Dec 2019 by Guizhou Meteorological Bureau.

Several preprocessing steps are conducted for preparing the dataset of our experiment. First, the precipitation data is accumulated over 10 minutes which still constitutes a reasonably high temporal resolution. To obtain a gridded dataset, the observations are then interpolated bilinearly onto a regular, spherical grid. The target grid comprises $w \times h = 48 \times 40$ data points in zonal and meridional direction, respectively, and covers a domain from $103.625°\text{E}$ to $109.5°\text{E}$ and $24.625°\text{N}$ to $29.5°\text{N}$ with $0.125°$ resolution. To obtain the data needed for training our CLGAN and the baseline models (see Sec. 2 and 3.1.4), we generate sliding sequences of 24 consecutive gridded data samples (frames) which comprise a temporal period of 240 minutes. 120 minutes (12 frames) of each sequence serve as input to predict the next 120 minutes (12 frames). Since there are many periods with no or only weak precipitation, we furthermore only select sequences whose averaged precipitation rate exceeds the empirical 60%-quantile of the complete dataset. This results in 35054 sequences for the subsampled dataset. Finally, a log-transformation is applied to each sequence to make the data more Gaussian-like. The log-transformation reads as follows: $x' = \ln(x+\varepsilon) - \ln(\varepsilon)$, where ε is a small constant (here 0.01). We use the data from 2015 to 2017 for training, the data of 2018 for validating, and the data of 2019 for testing.

3.3 Verification methods

As pointed out in Schultz et al. (2021) and more specifically for precipitation in Leinonen et al. (2020), precipitation nowcasting should be evaluated in terms of application-specific scores. This is due to the unique statistical properties of precipitation rates,

as well as the chaotic atmospheric processes which underpin the formation of precipitation. Additionally, we would like to emphasize that a single score alone can barely evaluate the model performance applied to high-dimensional data (Wilks, 2011). Therefore, we take several evaluation metrics into account to provide a comprehensive overview.

220   The first considered family of evaluation metrics is established for continuous quantities in the meteorological community. The root mean square error (RMSE) measures the distance between the predicted and the observed field on a grid-point level. The correlation coefficient (CC) measures the association or the linear relationship between the two fields. A perfect correlation would result in $CC = 1$, while $CC = 0$ indicates no linear relationship between forecast and observation on grid-point level.

The second set of scores is built on dichotomous events which are obtained by thresholding the gridded precipitation fields.
225   A 2×2 contingency table is commonly used to show the frequency of "yes" and "no" forecasts and occurrences and give a joint distribution for events with a precipitation rate exceeding a given threshold t_{pr}. According to the elements in the contingency table, a variety of categorical statistics can be computed to evaluate the dichotomous forecasts in particular aspects. Critical Success Index (CSI), also known as Threat Score, measures the fraction of hits with respect to the number of occurrences where the events are either forecasted or observed. The frequently applied Equitable Threat Score (ETS) is a variant of the CSI
230   and explicitly accounts for random forecasts which perform well just by chance (Wilks, 2011).

However, due to the highly non-linear and complex processes causing precipitation formation, scores acting on grid-point level are prone to penalize predictions which recover the high spatial variability, but fail to match exactly the observed precipitation field. The issue leads to the double penalty problem where the model gets penalized twice, once for missing the exact placement of a precipitation event and once for shifting it spatially (Ebert, 2008). To relax the requirement for exact
235   spatial matching, the Fractions Skill Score (FSS) is computed here as a fuzzy verification metric (Roberts, 2008). Similar to the CSI and ETS, the FSS operates on dichotomous events, but allows for spatial shifts by considering a local neighborhood around each grid point. Within this neighborhood, the fractional coverage of the precipitation events is calculated for both, the predictions and the observations. Let $f(m_i^s)$ and $f(o_i^s)$ denote the fraction of event grid boxes within the local neighborhood of size s around the grid point i in the model prediction and observation. The Fractions Brier Score (FBS) is given by:

$$FBS = \frac{1}{N} \sum_{i=1}^{N} (f(m_i^s) - f(o_i^s))^2. \tag{4}$$

which quantifies the quadratic difference between the prediction and the observation for all N grid points over the domain. The final FSS is then obtained with:

$$FSS = 1 - FBS/FBS_{worst}. \tag{5}$$

where FBS_{worst} is the sum of the squared fractions of events in the prediction and in the observation. Higher FSS-values
245   indicate better forecast, while it can be shown that a forecast becomes 'useful' when $FSS \geq 0.5$ is attained for a given neighborhood scale s (typically expressed in terms of squares with an edge length of N grid points).

Nonetheless, FSS also does not capture the spatial precipitation patterns since each grid point in the neighborhood is treated equally and no check for spatial coherence is undertaken. Thus, we additionally perform an object-based diagnostic evaluation, called MODE (Johnson and Wang, 2012; Johnson et al., 2013; Ji et al., 2020), to focus on pattern attributes such as location,

area and shape. To obtain the desired attributes, a convolutional filter of size k is first applied over the precipitation field. Afterwards, objects are defined by applying an threshold on the precipitation rate t_{Pr} and on the object area t_A. A fuzzy logic scheme is then used to merge and pair precipitation objects in the predicted and observed precipitation field. Finally, the object-based threat score (OTS; Johnson and Wang, 2012) is computed to verify how well the predicted precipitation patterns match the observed ones. Here, we choose the object area, the centroid location and the object shape (aspect ratio and orientation angle) as target attributes for computing the OTS.

All the mentioned verification methods are listed in Table 1 with a brief description. The details can be found in the corresponding references.

Table 1. Summary of the verification methods used in the paper

Verification method	Description	Formula or Reference	Notes
Root Mean Square Error	the average magnitude of the forecast errors	$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - Y_i')^2}$	$[0, +\infty)$;
Correlation Coefficient	the correspondence between the forecast and observed values	$\frac{\sum_{i=1}^{N}(Y_i - \bar{Y_i})(Y_i' - \bar{Y_i'})}{\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y_i})^2}\sqrt{\sum_{i=1}^{N}(Y_i' - \bar{Y_i'})^2}}$	$[-1, 1]$;
Critical Success Index	the correspondence between the forecast "yes" events and observed "yes" events	$\frac{hits}{hits + misses + false\ alarms}$	$[0, 1]$, 0 means no skills;
Equitable Threat Score	the correspondence between the forecast "yes" events and observed "yes" events (accounting for hits due to chance)	$\frac{hits - hits_{random}}{hits + false\ alarms}$ $hits_{random} =$ $\frac{(hits + misses)(hits + false\ alarms)}{total}$	$[-1/3, 1]$, 0 means no skills;
Fractions Skill Score	the spatial scales at which the forecast resembles the observations	(Roberts and Lean, 2008)	$[0,1]$, the smallest window size for which FSS$\geq$0.5 can be considered as "skillful scale";
Object-based Threat Score	the similarity between the forecast objects with the observed ones according to a series of attributes	(Davis et al., 2006)	$[0, 1]$, 0 means complete mismatch and 1 means perfect match;

To ease the comparison between the baseline models and the simplistic persistence forecast, we furthermore calculate skill scores (except for the FSS). In general, a skill score SS can be constructed by considering the target score S_m of the model, the score obtained with the reference forecast S_{ref} and the perfect score S_{perf}:

$$SS = \frac{S_m - S_{ref}}{S_{perf} - S_{ref}}. \tag{6}$$

The higher SS is, the better the model performs against the reference score. Perfect models thereby obtain $SS = 1$, while inferior models show up with $-\infty < SS < 0$. Note, that $S_{perf} = 0$ holds for the RMSE, whereas the other scores under consideration attain $S_{perf} = 1$. Since the size of our dataset is not unlimited, we also apply a block bootstrapping procedure to estimate sampling uncertainty (Efron and Tibshirani, 1994). The block bootstrapping procedure accounts for autocorrelation between the sliding sequences and thus, divides the dataset into non-overlapping blocks before resampling of the blocks with replacement is performed. Here, we set the block length to 10 hours (60 frames) and perform 1000 block bootstrapping steps.

4 Results

4.1 Quantitative evaluation

4.1.1 Point-wise evaluation metrics

In Fig. 3a-d, our model is compared to the baseline models in terms of the skill scores for the grid-point level evaluation metrics (CC, RMSE, CSI, and ETS). The skill scores are calculated by defining the Eulerian persistence as reference forecast. It is seen that the deep learning models (i.e. ConvLSTM, PredRNN-v2 and CLGAN) as well as the optical flow model DenseRotation outperform the persistence forecast after 20 minutes lead time in terms of the continuous scores (CC and RMSE). Among the video prediction models, PredRNN-v2 is superior over the others after the first 20 minutes, while ConvLSTM performs best for the longer lead times. CLGAN is not so competitive for RMSE and CC as PredRNN-v2 and ConvLSTM, while still outperforms the traditional optical flow model DenseRotation. It's noted that the Eulerian persistence performs well in the first 10 minutes. One possible reason why these complex models can barely beat the persistence forecast in the first lead step is that the precipitation systems are relatively invariant within this very short time period. In our case, the Eulerian persistence forecast is the latest observations available, hence, which is highly correlated to the ground truth at short lead times. With increasing lead times, its performance degrades quickly.

The comparisons in terms of the dichotomous scores (CSI and ETS) are given in Fig. 3c∼d. They demonstrate that CLGAN is superior to the other competitors at all the lead times for simulating heavy precipitation events (the threshold t_{Pr} is set to 8 mm/h here). The optical flow model DenseRotation performs well in the first 40 lead minutes while its skill scores decrease rapidly afterwards. By contrast, the advanced deep learning model PredRNN-v2 shows more potential for longer lead times. Although ConvLSTM outperforms on the continuous scores, it can barely capture the heavy precipitation events. A large performance degradation for the ConvLSTM is diagnosed at a lead time of 20 minutes. One reason of the difference is that the model performance is evaluated with the skill scores, which is affected by the choice of the reference model (here is the Eulerian persistence). For the first time step (lead time of 10 minutes), both ConvLSTM and Eulerian persistence can capture strong precipitation events and ConvLSTM is even better. However, ConvLSTM models are prone to produce blurry predictions in an autoregressive prediction task, where the errors in the prior forecasts are inherited to the later ones. Hence, the ConvLSTM model gets less efficient in the next few lead steps, while the Eulerian persistence performs fairly well. For

longer lead times, the performance of the Eulerian persistence forecasts quickly degrades and ConvLSTM again outperforms the persistence model with positive skill scores.

295   The comparisons of the model performance show that CLGAN is superior in terms of scores for dichotomous forecasts (CSI and ETS), while it is less competitive in terms of RMSE. This is due to the fact that our CLGAN encourages the model to generate forecasts which have a similar distribution as the ground truth data, rather than just reducing the averaged point-wise loss. Hence, more heavy precipitation events are predicted by the CLGAN model which improves the dichotomous forecast scores. However, more predicted high-value precipitation could cause larger biases, compared to the models only generating

300   low-value forecasts. The problem is magnified with the use of the point-by-point scores, i.e. the RMSE, which suffers from the double penalty issue.

4.1.2   Spatial verification scores

To further investigate the model performances, we now turn our attention to the spatial verification scores, the FSS and the MODE framework. Fig. 4a shows the model performance in terms of the FSS for a lead time of 60 minutes by the persistence

305   model (used as the reference forecast). The FSS is computed based on different neighbourhood sizes and thresholds of hourly precipitation rates. Specifically, the neighbourhood scale s in kilometers varies along the x-axis. The FSS-values for s attaining values of approximately 41, 69, 96, 152, and 290 kilometers (square boxes of 3, 5, 7, 11 and 21 grid points), respectively, are plotted and marked as box-whisker plots for varying precipitation thresholds. The boxes show the range of the first quartile (upper) to the third quartile (bottom) of the scores, and the whiskers are respectively the 95th percentile (upper) and 5th

310   percentile (bottom). As the threshold increases, FSS decreases, indicating that the persistence forecasts become increasingly imprecise for stronger precipitation events with a given spatial scale. For precipitation events exceeding $t_{Pr} = 8\,\text{mm/h}$, the persistence forecasts is considered useful ($FSS \geq 0.5$; see e.g. Roberts, 2008) for a neighborhood scale of $s \approx 69\,\text{km}$. Thus, the spatial accuracy of capturing these events is already fairly degraded and the neighborhood scale of $s \approx 69\,\text{km}$ (5 grid points) is applied to compute the FSS for other model in the following. Fig. 4b compares the models' performance in terms of the FSS

315   for heavy precipitation forecasting against the Eulerain persistence by illustrating the difference $\Delta FSS = FSS_i - FSS_{ref}$. Here, FSS_{ref} denotes the reference persistence model, whereas i is used to denote the other competing models. It is seen that all baseline models except from the ConvLSTM model can remarkably improve the spatial forecasting of such events, especially for longer lead times. Among them, CLGAN is superior to the others at all the lead times. DenseRotation performs well in the first lead hour, while PredRNN-v2 is promising for the further lead times.

320   ### 4.1.3   Object-based diagnostic evaluation

To fully access the performance of the models in predicting spatial precipitation attributes, i.e. area, location and shape, the MODE verification framework is applied. In the following, we present conditional quantile plots for the object area, for the location of the object centroid in east-west and north-south direction, for the aspect ratio and for the orientation angle of the precipitation objects to show more details of predicted and observed precipitation objects. These plots visualize the joint

325   distribution of the predictions and forecasts in a compact manner by applying a factorization into a conditional and marginal

distribution (Murphy and Winkler, 1987; Wilks, 2011). Fig 5 illustrates the joint distribution in terms of the likelihood-base rate factorization. While the solid lines illustrate the forecasts conditioned on the observations for all models, the marginal distribution of the observations is plotted as a histogram. Fig 5a shows the number of grid points of the observed and predicted precipitation objects, which represents the area of precipitation objects. It can be seen that CLGAN and PredRNN-v2 are able to capture object area fairly well. Only objects consisting of 90 to 130 grid points ($\sim 16,000 km^2$) are slightly underestimated (see Fig. 5a). However, the other competing models perform remarkably worse. Fig 5b~c show the distance between the centroid of the precipitation object and the western boundary and the southern boundary, which is again measured by the number of grid points. It is seen that the location of object centroids is generally well captured by all models. Stronger deviations are visible near the lateral boundaries, especially in east-west direction. The aspect ratio and the orientation angle are used to assess the predicted precipitation shape in Fig. 5d~e. Here, the aspect ratio is the ratio of the shorter to the longer edge of the precipitation objects. The orientation angle constitutes the angle between the precipitation objects and positive x-axis. CLGAN shows slight improvements over the other models where the central parts of the orientation angle and aspect ratio are well calibrated. However, larger deviations from the 1:1 reference line are obtained near the tails of the conditional distributions. This indicates that further work on simulation of precipitation shape is required.

4.2   Case study

To gain further insight into the realism of our predictions, a heavy precipitation event occurring on June 12th 2019 is visualized as an example (see Fig. 6) to compare the model performance with an 'eyeball' analysis. Fig. 6a shows the observed precipitation rates in mm/(10 min) for every 20 minutes over the forecast period starting at 06:50 BJT. It is seen that a fairly strong precipitation system moves from west to east while it further intensifies. The predictions of the different models are presented as difference plots in Fig. 6b~f. For the first 60 minutes, persistence and DenseRotation show up with the smallest discrepancies. However, for longer lead times, clear dipole structures in the difference plots indicate that the movement of the system is not captured. Thus, the Lagrangian persistence framework is inaccurate for longer lead times and more advanced models are required to capture the long-term dependence.

While the deep learning models also show up with increasing differences with longer lead times, they perform better in capturing the movement and the intensification of the precipitation system (see Fig. 6d~f). PredRNN-v2 tends to overestimate the precipitation intensity which causes large coherent areas of positive differences. The averaged RMSE over the study area of the PredRNN-v2 forecasts at the lead time of 2 hours is around 0.37 mm. ConvLSTM and CLGAN perform even better with smaller discrepancies, with a lower RMSE of 0.33 mm and 0.34 mm respectively. Consistent with the quantitative evaluation results, ConvLSTM outperforms the CLGAN model in terms of RMSE in this given case, whereas CLGAN obtains a higher CSI and ETS. The CSI of the CLGAN forecasts with a threshold of $8\,mm/h$ at the 2-hour lead time is around 0.11, 0.01 higher than the ConvLSTM model. The results demonstrate that the prediction by ConvLSTM has a lower bias, while CLGAN can capture more heavy precipitation grid-points. Compared to ConvLSTM, the difference plot of CLGAN contains more fine structures which indicates that CLGAN can generate more details of the precipitation system.

4.3 Ablation study

360 As shown in Eq. 3, the loss function used in our CLGAN model consists of two terms: the adversarial loss $\mathcal{L}^{GAN}$ and the reconstruction loss $\mathcal{L}^2$. To assess the contribution of each loss term on the forecasts performance, sensitivity experiments on the weight of the reconstruction loss in CLGAN were carried out. Larger weight of the adversarial loss (= smaller weight of reconstruction loss) is equivalent to a stronger contribution made by the GAN component. Fig. 7 presents the results of the CLGAN model with different weights λ assigned to the $\mathcal{L}^2$ loss. It is seen that the RMSE is increased when reducing the

365 weight of $\mathcal{L}^2$ loss (Fig. 7a). However, the scores for dichotomous events and fuzzy verification framework reveal improvements with for smaller λ. The model using a pure reconstruction loss ($\lambda = 1$ in Eq 3) performs significantly worse than the model applying an adversarial loss in terms of CSI and FSS (Fig. 7b∼c). Similar results are obtained in terms of the OTS (see Fig. 7d). The results of sensitivity experiments indicate that the adversarial training with the GAN-component encourages the model to generate forecasts which are more similar to the ground truth data. Despite a slight increase in RMSE, a relatively stronger

370 contribution of the GAN-component helps to capture the statistical properties of the observed precipitation (on the tail as well as their spatial attributes) which in turn improves the prediction of strong precipitation events.

5 Conclusion and Discussion

A novel architecture CLGAN is proposed in this work which leverages the merits of U-Net, ConvLSTM and GAN components, to generate high quality precipitation predictions up to 2 hours over Guizhou, China. The Eulerian persistence is used as

375 the reference model to compare against the conventional optical flow method DenseRotation, as well as two competing video prediction models (ConvLSTM and PredRNN-v2). A Guizhou AWS_ML precipitation dataset is set-up for the nowcasting task based on minute-level precipitation measurements of AWS gauges. The model performance is comprehensively evaluated by a series of domain-specific evaluation metrics, including point-by-point and object-based verification methods. The results demonstrate that DL-based video prediction models are generally superior to the conventional methods, especially for the lead

380 times exceeding 60 minutes. However, the use of grid-point level losses (e.g. $\mathcal{L}^1$ or $\mathcal{L}^2$ loss) diminishes their capability in capturing heavy precipitation events. Since heavy precipitation events are strongly under-represented in the data during training, the models optimized solely on grid-point level losses favor to predict weak precipitation rates to avoid large error contributions from strong precipitation events (double penalty problem). By contrast, the GAN-component of CLGAN encourages the generator to create predictions that share the statistical properties of observed precipitation, which makes it superior to the

385 baseline and the competing models in dichotomous and spatial scores for heavy precipitation events.

Compared to the conventional methods, our results indicate that video prediction models with deep neural networks have a better capability of learning abstractions from data, which in turn can improve the prediction of complex evolving systems. By learning the statistical dependency within the continuous sequence of precipitation data, video prediction models can simulate the precipitation patterns up to 2 hours ahead fairly well. Since NWP models suffer from the spin-up issue in the first 6

390 hours and the conventional approaches fail to capture long-term dependency, video prediction models show potential as a promising and reliable way for precipitation nowcasting. However, a model performance degradation is expected for longer

lead time, i.e. after 2 hours, due to the error accumulation in the auto-regressive prediction task. The quick evolution of the convective precipitation systems is furthermore challenging. Our results also demonstrate that it is arduous to capture the shape of precipitation patterns by DL-based models as demonstrated by the MODE scores. Future work may try to integrate domain-specific evaluation metrics for spatial forecasts (e.g. FSS and MODE) as a loss function in DL-based models for precipitation nowcasting. Additionally, we also see that a trade-off exists between evaluations on grid-point and object-based level when the adversarial loss is varied. Grid search for the optimal combination of loss function coefficients is required to generate realistic forecasts with a low bias.

Beyond that, it's appealing to embed more predictors which could be retrieved from NWP models, e.g. the vertical velocity, water vapor, thermal and other environmental conditions. Literature shows that the corresponding predictors and physical constraints can greatly improve the simulation of the targeted variable (Daw et al., 2017; Gong et al., 2022). A careful selection of the predictors and an appropriate embedding solution are subject to our future work. In addition, GAN models can easily be adapted to a probabilistic framework. By adding noise as an additional input, ensemble forecasts can be obtained from which a quantification of the forecast uncertainty can be deduced (Mordido et al., 2018). A probabilistic nowcasting system is appealing due to the strong inherent uncertainties in the dynamics of precipitation patterns. It's furthermore noted that ensemble predictions corresponding to several future realisations provide the possibility to issue more strong precipitation events (cf. Ravuri et al. (2021)). While this study focused on assessing the need for an adversarial loss formulation, future work will be directed towards a probabilistic nowcasting system.

Code and data availability. The Guizhou AWS_ML precipitation dataset and the exact version of the video prediction models used in this paper are archived on Zenodo (https://zenodo.org/record/7278016#.Y2ppyL7MIuQ). A frozen code repository can be obtained here: https://gitlab.jsc.fz-juelich.de/esde/machine-learning/ambs/-/tree/ambs_gmd_nowcasting_v1.0. The dataset and scripts can help users to reproduce the results on their local machines or high-performance computers. By using these data and models, it is highly recommended to follow the README.md file of the code repository to run the end-to-end workflow.

Author contributions. Yan Ji, Bing Gong and Michael Langguth equally contributed to this work. Yan Ji and Bing Gong contributed to the method development and perform the experiments. Yan Ji wrote the manuscript draft and all authors reviewed and edited the manuscript in several iterations.

Competing interests. The authors declare that they have no conflict of interest

Acknowledgements. The authors acknowledge funding from the DeepRain project under grant agreement 01 IS18047A from the Bundesministerium für Bildung und Forschung (BMBF), from the European Union H2020 MAELSTROM project (grant No. 955513, co-funding by

420 BMBF), and from the ERC Advanced grant IntelliAQ (grant no. 787576). We thank Dexuan Kong for preparing the datasets used in our research, as well as Martin G. Schultz for the helpful scientific discussions.
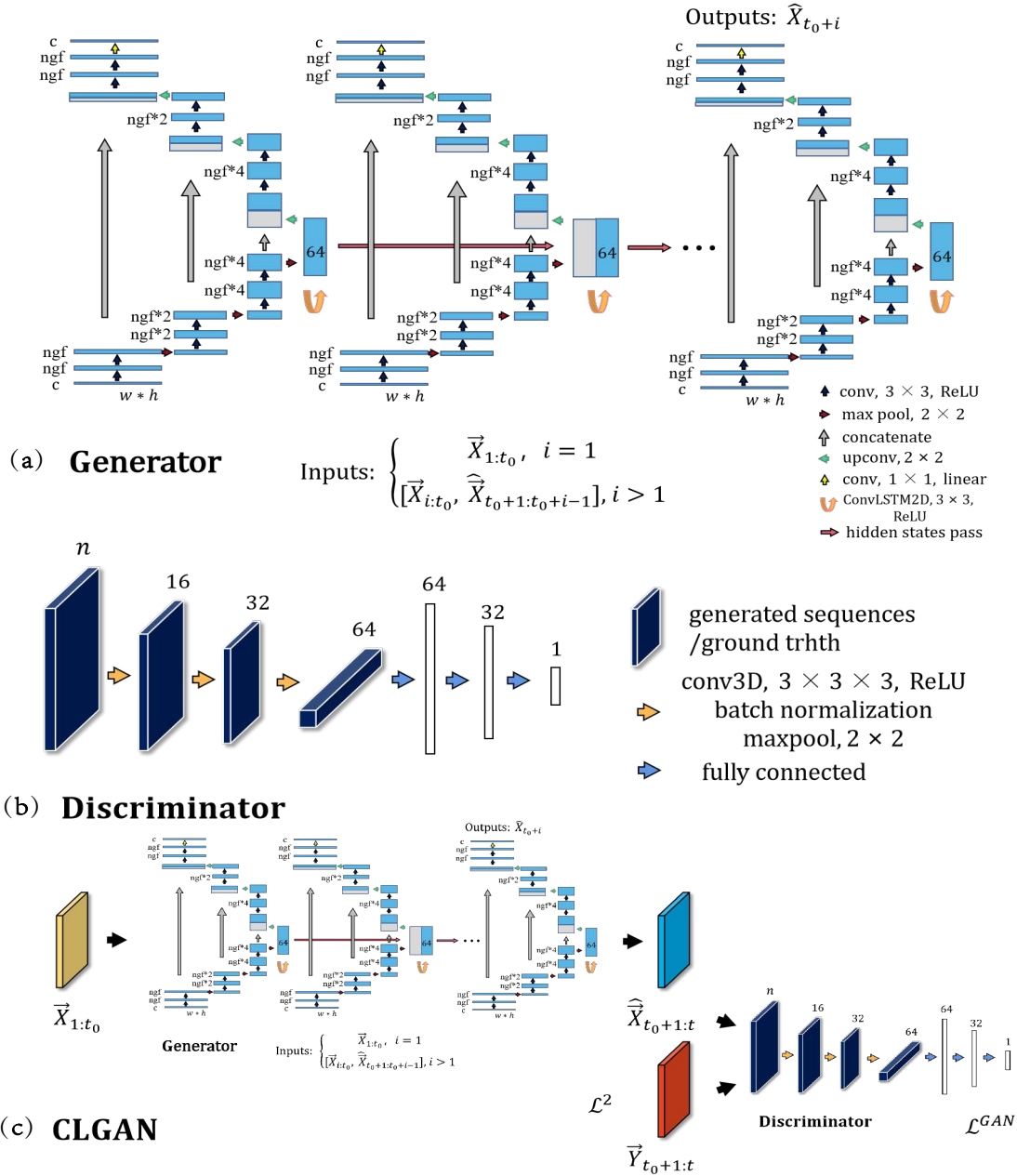
Figure 1. The details of the proposed CLGAN model; (a) Generator: The illustration is presented for a given forecast step i. If $i = 1$, the inputs are observed sequences $\vec{X}_{1:t_0}$. Otherwise the inputs are combined sequences of observed ones $\vec{X}_{i:t_0}$, and predicted ones $\widehat{\vec{X}}_{t_0+1:t_0+i-1}$. The output is the model prediction $\widehat{\vec{X}}_{t_0+i}$. c is the number of channels of inputs, here is 1. ngf denotes the number of filters in the first layer of U-Net. (b) Discriminator: n is the length of output sequences. (c) CLGAN: The outputs of generator are the predicted sequences $\widehat{\vec{X}}_{t_0+1:t}$. $\vec{Y}_{t_0+1:t}$ constitutes the corresponding ground-truth. $\mathcal{L}^2$ and $\mathcal{L}^{GAN}$ are the reconstruction loss and adversarial loss, respectively.
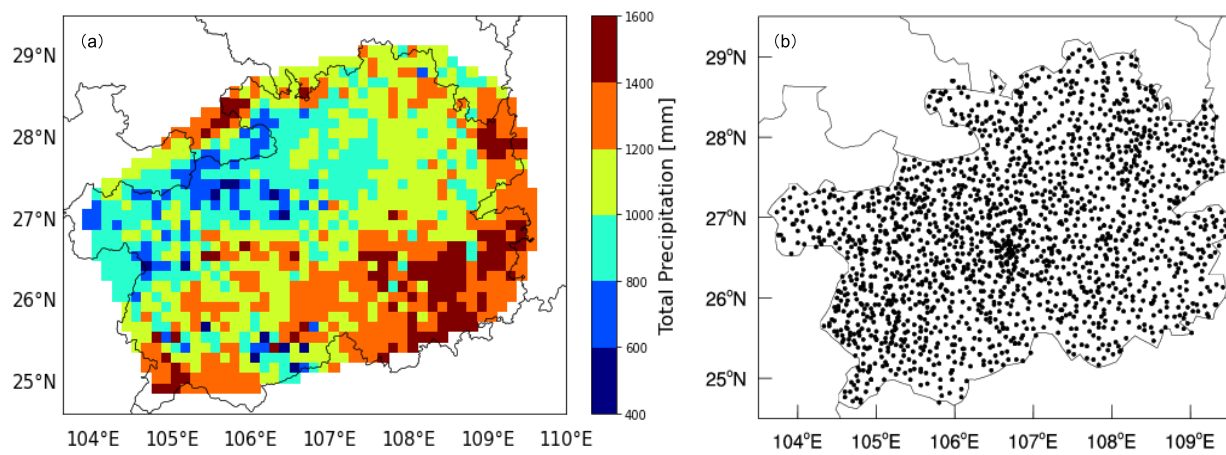
Figure 2. (a) Annual average cumulative precipitation in Guizhou from 2015 to 2019; (b) The spatial distribution of AWS over Guizhou.
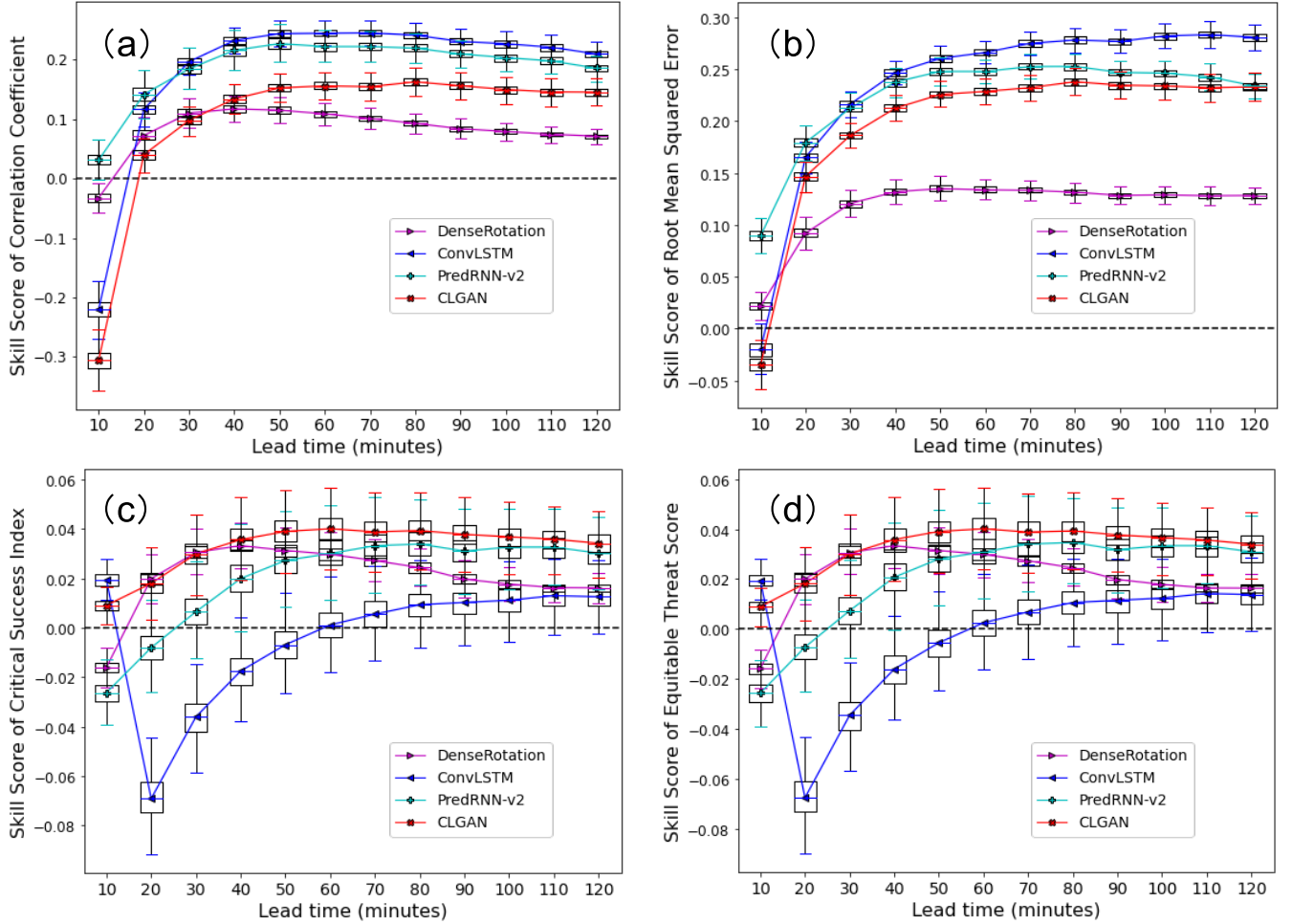
Figure 3. Box-whisker plots for skill scores of (a) CC, (b) RMSE, (c) CSI and (d) ETS averaged over the testing period with the Eulerian persistence as the reference forecast. The boxes show the range of the first quartile (upper) to the third quartile (bottom) of the skill scores, and the whiskers denote the 95th percentile (upper) and 5th percentile (bottom), respectively. The threshold t_{pr} of CSI and ETS is 8 mm/h.
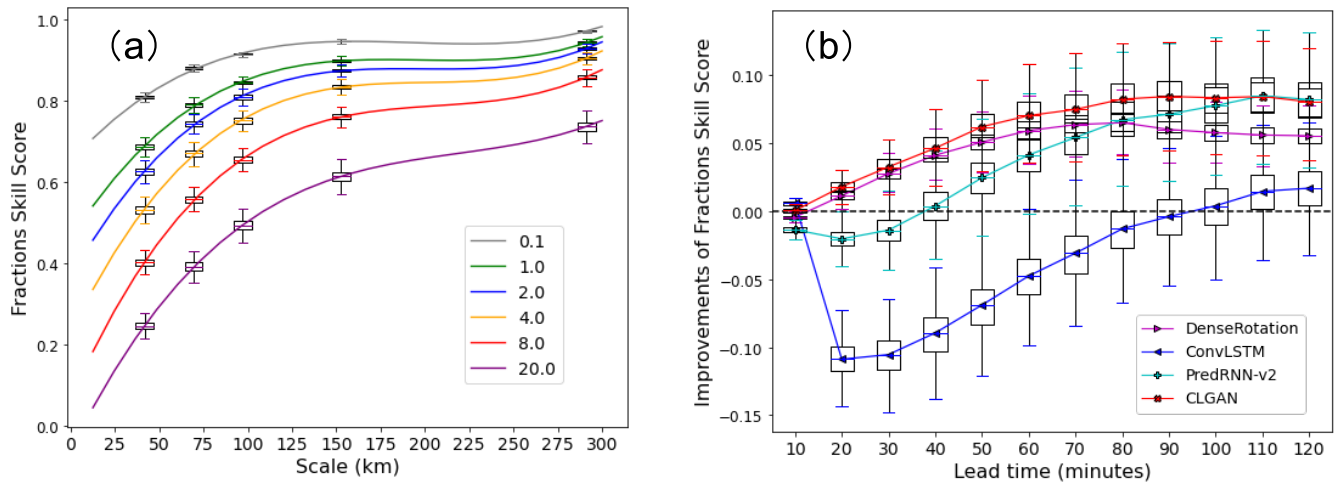
Figure 4. (a) FSS of persistence for different scales and intensity thresholds at 60 lead minutes for precipitation nowcasting. (b) Improvements of different models compared with persistence in terms of FSS with a threshold t_{pr} of 8 mm/h and a neighbourhood size of 5. The boxes show the range of the first quartile (upper) to the third quartile (bottom) of the skill scores, and the whiskers denote the 95th percentile (upper) and 5th percentile (bottom), respectively.
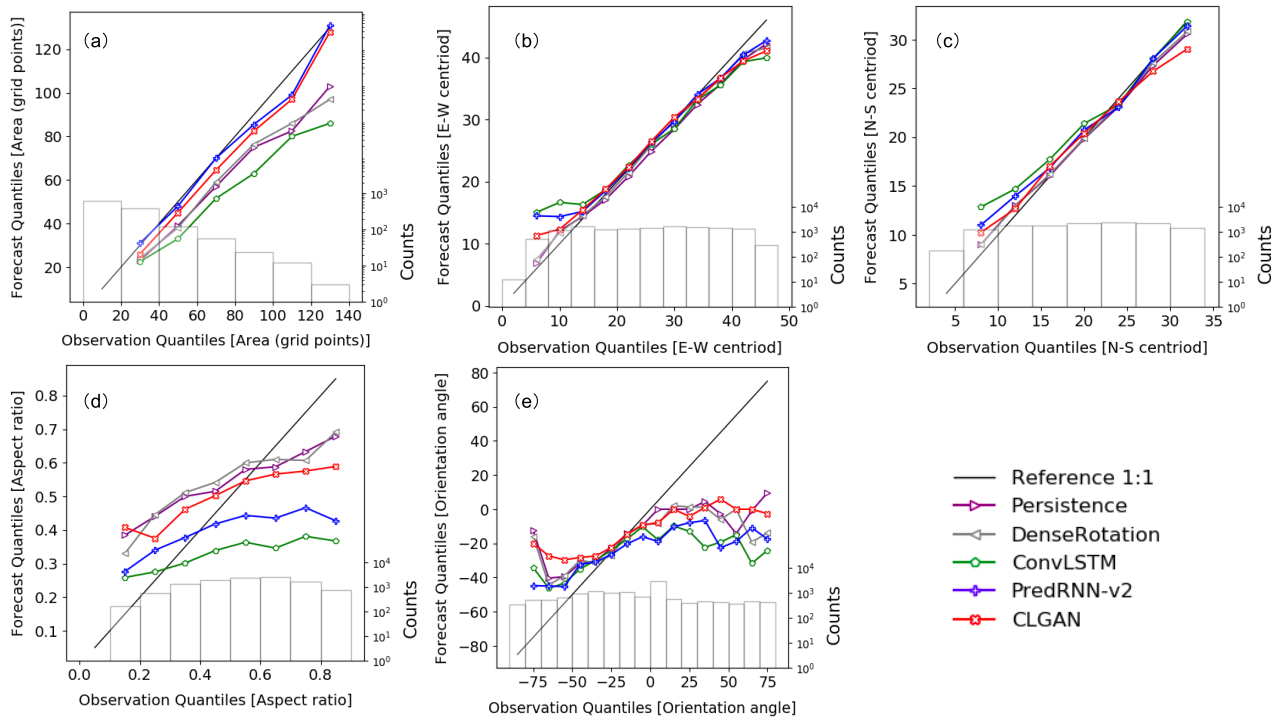
Figure 5. Conditional quantile plots in terms of the likelihood base-rate factorization for [(a) area, (b) E-W centroid and N-S centroid location, (d) aspect ratio, and (e) orientation angle, at the lead time of 60 minutes. The solid black line is the 1:1 reference line. The marginal distribution of the observations is presented as a histogram.
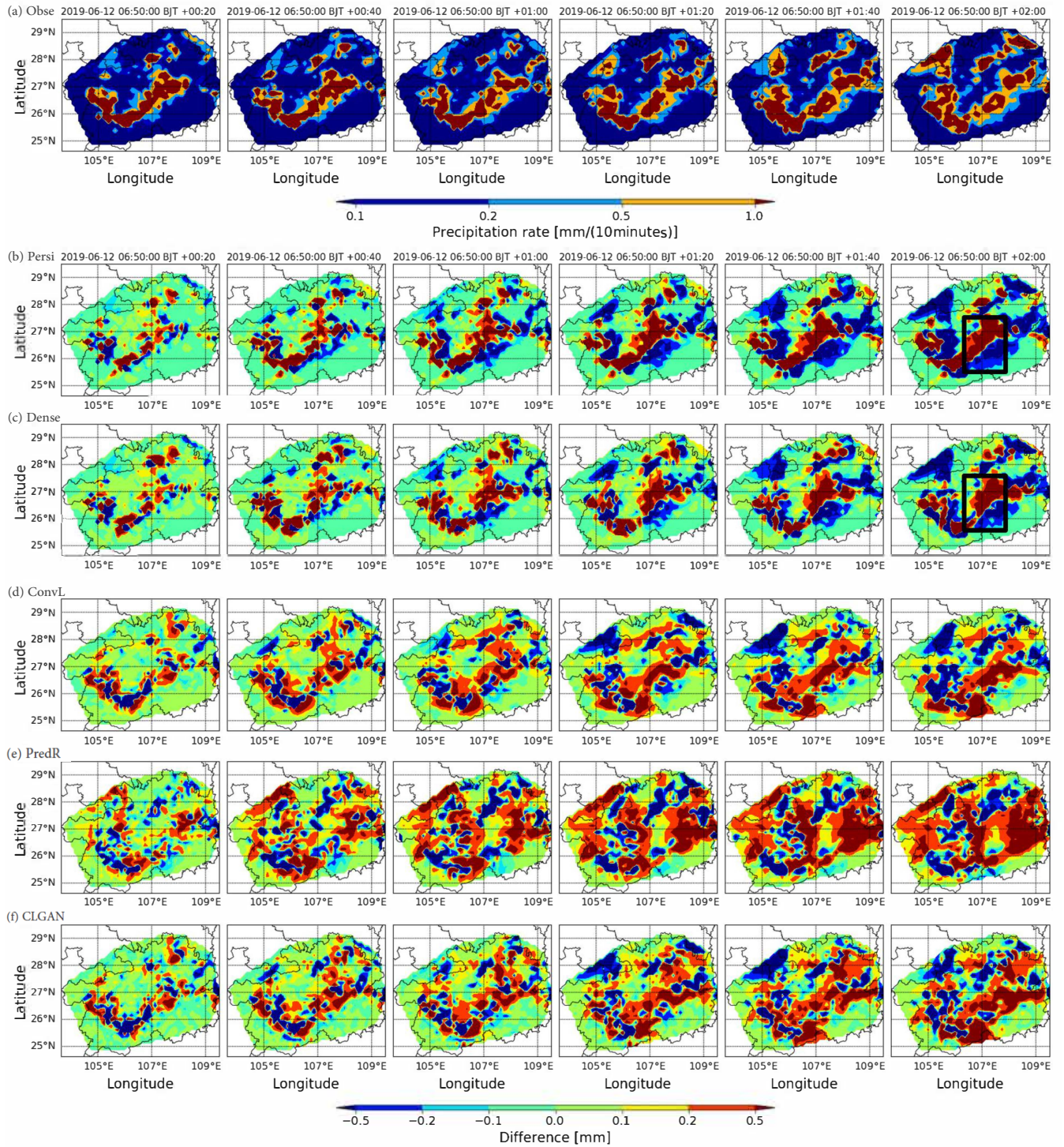
Figure 6. A case study for a rain system moving from west to east while intensifying (a) Observation (Obse). The predictions of all models are illustrated with difference plots: (b) Persistence (Persi), (c) DenseRotation (Dense), (d) ConvLSTM (ConvL), (e) PredRNN-v2 (PredR) and (f) CLGAN. The initial time of the prediction period is 06:50 on June 12th 2019 (BJT).
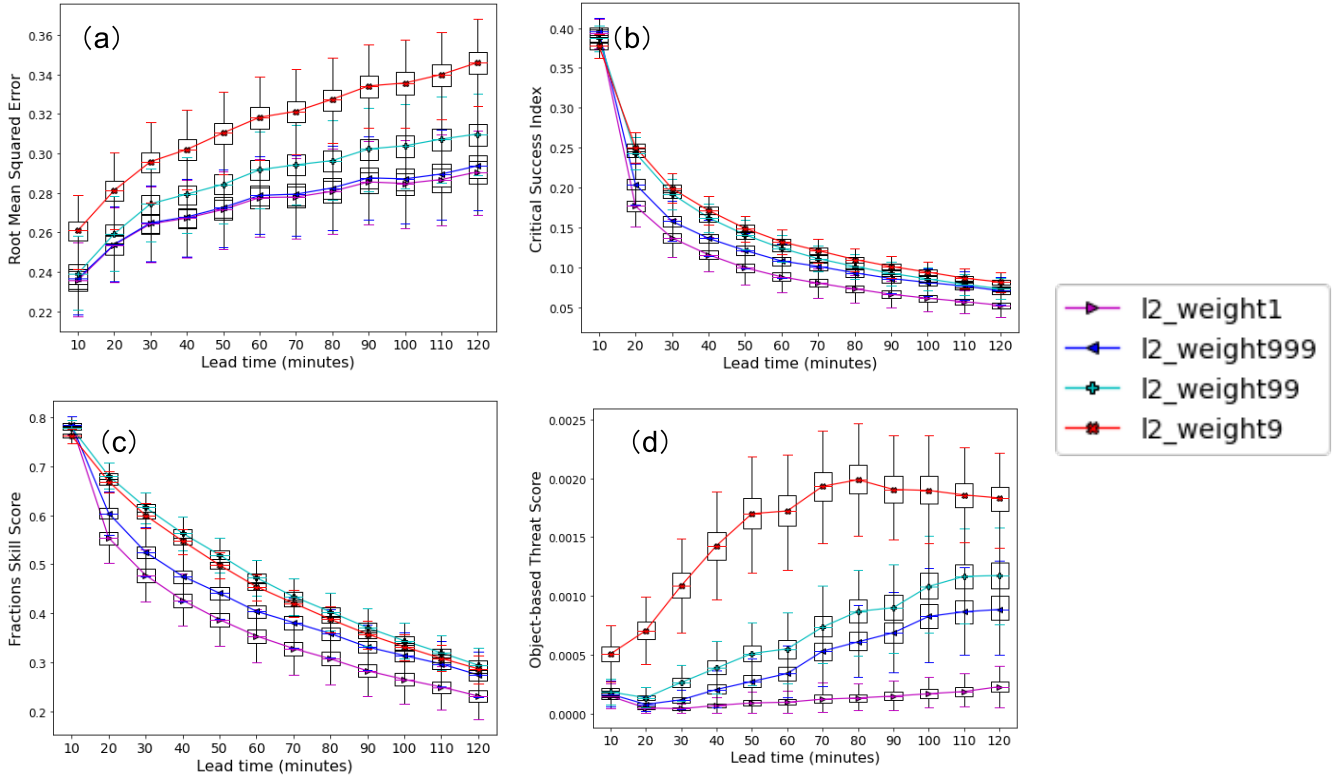
Figure 7. Mean scores [(a) RMSE, (b) CSI, (c) FSS and (d) OTS] for all lead times over the verification period for different weights of the $\mathcal{L}^2$ loss in CLGAN. The terms 'weight1', 'weight999', 'weight99' and 'weight9' denote the weights of the $\mathcal{L}^2$ loss component λ with the corresponding values of 1, 0.999, 0.99 and 0.9, respectively. As in Fig. 3, $t_{Pr} = 8\,\text{mm/h}$ is chosen for CSI, OTS and FSS together with setting s to 5 grid points for the latter. The area threshold t_A of OTS is set to 9 grid points.

References

Austin, G. and Bellon, A.: The use of digital weather radar records for short-term precipitation forecasting, Quarterly Journal of the Royal Meteorological Society, 100, 658–664, https://doi.org/http://dx.doi.org/10.1002/qj.49710042612, 1974.

Ayzel, G., Heistermann, M., and Winterrath, T.: Optical flow models as an open benchmark for radar-based precipitation nowcasting (rainymotion v0. 1), Geoscientific Model Development, 12, 1387–1402, https://doi.org/http://dx.doi.org/10.5194/gmd-12-1387-2019, 2019.

Ayzel, G., Scheffer, T., and Heistermann, M.: RainNet v1. 0: a convolutional neural network for radar-based precipitation nowcasting, Geoscientific Model Development, 13, 2631–2644, https://doi.org/http://dx.doi.org/10.5194/gmd-13-2631-2020, 2020.

Bowler, N. E., Pierce, C. E., and Seed, A.: Development of a precipitation nowcasting algorithm based upon optical flow techniques, Journal of Hydrology, 288, 74–91, https://doi.org/http://dx.doi.org/10.1016/j.jhydrol.2003.11.011, 2004.

Bowler, N. E., Pierce, C. E., and Seed, A. W.: STEPS: A probabilistic precipitation forecasting scheme which merges an extrapolation nowcast with downscaled NWP, Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography, 132, 2127–2155, https://doi.org/http://dx.doi.org/10.1256/qj.04.100, 2006.

Davis, C., Brown, B., and Bullock, R.: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas, Monthly Weather Review, 134, 1772–1784, https://doi.org/http://dx.doi.org/10.1175/MWR3145.1, 2006.

Daw, A., Karpatne, A., Watkins, W. D., Read, J. S., and Kumar, V.: Physics-guided neural networks (pgnn): An application in lake temperature modeling, in: Knowledge-Guided Machine Learning, pp. 353–372, Chapman and Hall/CRC, https://doi.org/http://dx.doi.org/10.1201/9781003143376-15, 2017.

Dixon, M. and Wiener, G.: TITAN: Thunderstorm identification, tracking, analysis, and nowcasting—A radar-based methodology, Journal of atmospheric and oceanic technology, 10, 785–797, https://doi.org/http://dx.doi.org/10.1175/1520-0426(1993)010%3C0785:TTITAA%3E2.0.CO;2, 1993.

Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S., and Pal, C.: The importance of skip connections in biomedical image segmentation, in: Deep learning and data labeling for medical applications, pp. 179–187, Springer, https://doi.org/http://dx.doi.org/10.1007/978-3-319-46976-8_19, 2016.

Ebert, E. E.: Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework, Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling, 15, 51–64, https://doi.org/http://dx.doi.org/10.1002/met.25, 2008.

Ebert, F., Finn, C., Lee, A. X., and Levine, S.: Self-Supervised Visual Planning with Temporal Skip Connections, in: CoRL, pp. 344–356, 2017.

Efron, B. and Tibshirani, R. J.: An introduction to the bootstrap, CRC press, https://doi.org/http://dx.doi.org/10.1201/9780429246593, 1994.

Ganguly, A. R. and Bras, R. L.: Distributed quantitative precipitation forecasting using information from radar and numerical weather prediction models, Journal of Hydrometeorology, 4, 1168–1180, https://doi.org/http://dx.doi.org/10.1175/1525-7541(2003)004%3C1168:DQPFUI%3E2.0.CO;2, 2003.

Garcia-Garcia, A., Martinez-Gonzalez, P., Oprea, S., Castro-Vargas, J. A., Orts-Escolano, S., Garcia-Rodriguez, J., and Jover-Alvarez, A.: The robotrix: An extremely photorealistic and very-large-scale indoor dataset of sequences with robot trajectories and interactions, in: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 6790–6797, IEEE, https://doi.org/http://dx.doi.org/10.1109/IROS.2018.8594495, 2018.

Germann, U. and Zawadzki, I.: Scale-dependence of the predictability of precipitation from continental radar images. Part I: Description of the methodology, Monthly Weather Review, 130, 2859–2873, https://doi.org/http://dx.doi.org/10.1175/1520-0493(2002)130%3C2859:SDOTPO%3E2.0.CO;2, 2002.

Gong, B., Langguth, M., Ji, Y., Mozaffari, A., Stadtler, S., Mache, K., and Schultz, M. G.: Temperature forecasting by deep learning methods, Geoscientific Model Development Discussions, pp. 1–35, https://doi.org/https://doi.org/10.5194/gmd-2021-430, 2022.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y.: Generative adversarial networks, Communications of the ACM, 63, 139–144, https://doi.org/http://dx.doi.org/10.1145/3422622, 2020.

Grecu, M. and Krajewski, W.: A large-sample investigation of statistical procedures for radar-based short-term quantitative precipitation forecasting, Journal of hydrology, 239, 69–84, https://doi.org/http://dx.doi.org/10.1016/S0022-1694(00)00360-7, 2000.

Harris, L., McRae, A. T., Chantry, M., Dueben, P. D., and Palmer, T. N.: A Generative Deep Learning Approach to Stochastic Downscaling of Precipitation Forecasts, arXiv preprint arXiv:2204.02028, https://doi.org/http://dx.doi.org/10.1029/2022MS003120, 2022.

Hu, A., Cotter, F., Mohan, N., Gurau, C., and Kendall, A.: Probabilistic future prediction for video scene understanding, in: European Conference on Computer Vision, pp. 767–785, Springer, https://doi.org/http://dx.doi.org/10.1007/978-3-030-58517-4_45, 2020.

Ji, L., Zhi, X., Simmer, C., Zhu, S., and Ji, Y.: Multimodel ensemble forecasts of precipitation based on an object-based diagnostic evaluation, Monthly Weather Review, 148, 2591–2606, https://doi.org/http://dx.doi.org/10.1175/MWR-D-19-0266.1, 2020.

Johnson, A. and Wang, X.: Object-based evaluation of a storm-scale ensemble during the 2009 NOAA Hazardous Weather Testbed Spring Experiment, Monthly weather review, 141, 1079–1098, https://doi.org/http://dx.doi.org/10.1175/MWR-D-12-00140.1, 2012.

Johnson, A., Wang, X., Kong, F., and Xue, M.: Object-based evaluation of the impact of horizontal grid spacing on convection-allowing forecasts, Monthly weather review, 141, 3413–3425, https://doi.org/http://dx.doi.org/10.1175/MWR-D-13-00027.1, 2013.

Johnson, J., MacKeen, P. L., Witt, A., Mitchell, E. D. W., Stumpf, G. J., Eilts, M. D., and Thomas, K. W.: The storm cell identification and tracking algorithm: An enhanced WSR-88D algorithm, Weather and forecasting, 13, 263–276, https://doi.org/http://dx.doi.org/10.1175/1520-0434(1998)013%3C0263:TSCIAT%3E2.0.CO;2, 1998.

Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980, https://doi.org/https://doi.org/10.48550/arXiv.1412.6980, 2014.

Kroeger, T., Timofte, R., Dai, D., and Van Gool, L.: Fast optical flow using dense inverse search, in: European Conference on Computer Vision, pp. 471–488, Springer, https://doi.org/http://dx.doi.org/10.1007/978-3-319-46493-0_29, 2016.

Leinonen, J., Nerini, D., and Berne, A.: Stochastic super-resolution for downscaling time-evolving atmospheric fields with a generative adversarial network, IEEE Transactions on Geoscience and Remote Sensing, https://doi.org/http://dx.doi.org/10.1109/TGRS.2020.3032790, 2020.

Li, D., Liu, Y., and Chen, C.: MSDM v1. 0: A machine learning model for precipitation nowcasting over eastern China using multisource data, Geoscientific Model Development, 14, 4019–4034, https://doi.org/http://dx.doi.org/10.5194/gmd-14-4019-2021, 2021.

Liu, W., Luo, W., Lian, D., and Gao, S.: Future frame prediction for anomaly detection–a new baseline, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6536–6545, https://doi.org/http://dx.doi.org/10.1109/CVPR.2018.00684, 2018.

Mathieu, M., Couprie, C., and LeCun, Y.: Deep multi-scale video prediction beyond mean square error, arXiv preprint arXiv:1511.05440, https://doi.org/https://doi.org/10.48550/arXiv.1511.05440, 2015.

Matsunobu, T., Zarboo, A., Barthlott, C., and Keil, C.: Impact of combined microphysical uncertainties on convective clouds and precipitation in ICON-D2-EPS forecasts during different synoptic control, Weather and Climate Dynamics Discussions, pp. 1–25, https://doi.org/https://doi.org/10.5194/wcd-2022-17, 2022.

Mordido, G., Yang, H., and Meinel, C.: Dropout-gan: Learning from a dynamic ensemble of discriminators, arXiv preprint arXiv:1807.11346, https://doi.org/https://doi.org/10.48550/arXiv.1807.11346, 2018.

Murphy, A. H. and Winkler, R. L.: A general framework for forecast verification, Monthly weather review, 115, 1330–1338, https://doi.org/http://dx.doi.org/10.1175/1520-0493(1987)115%3C1330:AGFFFV%3E2.0.CO;2, 1987.

500 Oprea, S., Martinez-Gonzalez, P., Garcia-Garcia, A., Castro-Vargas, J. A., Orts-Escolano, S., Garcia-Rodriguez, J., and Argyros, A.: A review on deep learning techniques for video prediction, IEEE Transactions on Pattern Analysis and Machine Intelligence, https://doi.org/http://dx.doi.org/10.1109/TPAMI.2020.3045007, 2020.

Price, I. and Rasp, S.: Increasing the accuracy and resolution of precipitation forecasts using deep generative models, arXiv preprint arXiv:2203.12297, https://doi.org/https://doi.org/10.48550/arXiv.2203.12297, 2022.

505 Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., et al.: Skillful Precipitation Nowcasting using Deep Generative Models of Radar, arXiv preprint arXiv:2104.00954, https://doi.org/https://doi.org/10.1038/s41586-021-03854-z, 2021.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., et al.: Deep learning and process understanding for data-driven Earth system science, Nature, 566, 195–204, https://doi.org/http://dx.doi.org/10.1038/s41586-019-0912-1, 2019.

510 Rinehart, R. and Garvey, E.: Three-dimensional storm motion detection by conventional weather radar, Nature, 273, 287–289, https://doi.org/http://dx.doi.org/10.1038/273287a0, 1978.

Roberts, N.: Assessing the spatial and temporal variation in the skill of precipitation forecasts from an NWP model, Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling, 15, 163–169, https://doi.org/http://dx.doi.org/10.1002/met.57, 2008.

515 Roberts, N. M. and Lean, H. W.: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events, Monthly Weather Review, 136, 78–97, https://doi.org/https://doi.org/10.1175/2007MWR2123.1, 2008.

Ronneberger, O., Fischer, P., and Brox, T.: U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, pp. 234–241, Springer, https://doi.org/http://dx.doi.org/10.1007/978-3-319-24574-4_28, 2015.

520 Schultz, M., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L., Mozaffari, A., and Stadtler, S.: Can deep learning beat numerical weather prediction?, Philosophical Transactions of the Royal Society A, 379, 20200097, https://doi.org/http://dx.doi.org/10.1098/rsta.2020.0097, 2021.

Sha, Y., Gagne II, D. J., West, G., and Stull, R.: Deep-learning-based gridded downscaling of surface meteorological variables in complex terrain. Part II: Daily precipitation, Journal of Applied Meteorology and Climatology, 59, 2075–2092, 525 https://doi.org/http://dx.doi.org/10.1175/JAMC-D-20-0058.1, 2020.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., and Woo, W.-c.: Convolutional LSTM network: A machine learning approach for precipitation nowcasting, Advances in neural information processing systems, 28, 2015.

Shi, X., Gao, Z., Lausen, L., Wang, H., Yeung, D.-Y., Wong, W.-k., and Woo, W.-c.: Deep learning for precipitation nowcasting: A benchmark and a new model, arXiv preprint arXiv:1706.03458, https://doi.org/https://doi.org/10.48550/arXiv.1706.03458, 2017.

530 Sønderby, C. K., Espeholt, L., Heek, J., Dehghani, M., Oliver, A., Salimans, T., Agrawal, S., Hickey, J., and Kalchbrenner, N.: Metnet: A neural weather model for precipitation forecasting, arXiv preprint arXiv:2003.12140, https://doi.org/https://doi.org/10.48550/arXiv.2003.12140, 2020.

Sun, J., Xue, M., Wilson, J. W., Zawadzki, I., Ballard, S. P., Onvlee-Hooimeyer, J., Joe, P., Barker, D. M., Li, P.-W., Golding, B., et al.: Use of NWP for nowcasting convective precipitation: Recent progress and challenges, Bulletin of the American Meteorological Society, 95, 409–426, https://doi.org/http://dx.doi.org/10.1175/BAMS-D-11-00263.1, 2014.

Vasiloff, S. V., Seo, D.-J., Howard, K. W., Zhang, J., Kitzmiller, D. H., Mullusky, M. G., Krajewski, W. F., Brandes, E. A., Rabin, R. M., Berkowitz, D. S., et al.: Improving QPE and very short term QPF: An initiative for a community-wide integrated approach, Bulletin of the American Meteorological Society, 88, 1899–1911, https://doi.org/http://dx.doi.org/10.1175/BAMS-88-12-1899, 2007.

Wang, Y., Long, M., Wang, J., Gao, Z., and Philip, S. Y.: Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms, in: Advances in Neural Information Processing Systems, pp. 879–888, 2017.

Wang, Y., Wu, H., Zhang, J., Gao, Z., Wang, J., Yu, P. S., and Long, M.: PredRNN: A Recurrent Neural Network for Spatiotemporal Predictive Learning, arXiv preprint arXiv:2103.09504, https://doi.org/http://dx.doi.org/10.1109/TPAMI.2022.3165153, 2021.

Wilks, D. S.: Statistical methods in the atmospheric sciences, vol. 100, Academic press, 2011.

Wilson, J. W., Feng, Y., Chen, M., and Roberts, R. D.: Nowcasting challenges during the Beijing Olympics: Successes, failures, and implications for future nowcasting systems, Weather and Forecasting, 25, 1691–1714, https://doi.org/http://dx.doi.org/10.1175/2010WAF2222417.1, 2010.

Wolberg, G.: Digital image warping, vol. 10662, IEEE computer society press Los Alamitos, CA, 1990.

Woo, W.-c. and Wong, W.-k.: Operational application of optical flow techniques to radar-based rainfall nowcasting, Atmosphere, 8, 48, https://doi.org/http://dx.doi.org/10.3390/atmos8030048, 2017.

Xie, S., Wang, Y.-C., Lin, W., Ma, H.-Y., Tang, Q., Tang, S., Zheng, X., Golaz, J.-C., Zhang, G. J., and Zhang, M.: Improved diurnal cycle of precipitation in E3SM with a revised convective triggering function, Journal of Advances in Modeling Earth Systems, 11, 2290–2310, https://doi.org/http://dx.doi.org/10.1029/2019MS001702, 2019.

Zahraei, A., Hsu, K.-l., Sorooshian, S., Gourley, J., Lakshmanan, V., Hong, Y., and Bellerby, T.: Quantitative precipitation nowcasting: A Lagrangian pixel-based approach, Atmospheric Research, 118, 418–434, https://doi.org/http://dx.doi.org/10.1016/j.atmosres.2012.07.001, 2012.

Zahraei, A., Hsu, K.-l., Sorooshian, S., Gourley, J. J., Hong, Y., and Behrangi, A.: Short-term quantitative precipitation forecasting using an object-based approach, Journal of hydrology, 483, 1–15, https://doi.org/http://dx.doi.org/10.1016/j.jhydrol.2012.09.052, 2013.