Assessing U-Net Model Performance in Weather Radar-Based Precipitation Nowcasting: A Reflectivity Threshold Analysis

Rafael Rocha[1,2], **Ewerton Oliveira**[1,2], **Eduardo Carvalho**[1], **Renata Tedeschi**[1], **Claudia Costa**[1], **Douglas Ferreira**[1], **Ronnie Alves**[1]

[1]Instituto Tecnológico Vale
Belém – PA – Brazil

[2]Federal University of Pará
Belém – PA – Brazil

{rafael.lima.rocha, ewerton.oliveira}@pq.itv.org,
{eduardo.costa.carvalho, renata.tedeschi,
claudia.costa, douglas.silva.ferreira,
ronnie.alves}@itv.org

Abstract. *Severe weather events pose a global threat, causing property damage and endangering lives. Weather-related disasters account for 50% of all natural and technological disasters. The development of accurate prediction systems are crucial for early warnings and mitigation. The present study evaluates the effectiveness of the U-Net model in weather radar-based precipitation nowcasting, considering reflectivity thresholds. Visual comparison and evaluation metrics are used to assess observed and predicted reflectivity. The 10 dBZ threshold achieved a prominent result, accurately predicting over 75% of values above the reflectivity threshold. Results contribute to improving severe weather prediction and decision-making.*

1. Introduction

Severe weather phenomena present a significant challenge faced by many countries worldwide. Storms, and floods, for example, represent a threat to the safety and well-being of communities worldwide. These phenomena can result in heavy rainfall, strong winds, hail, and flash floods, causing property damage, disruption of essential services, and even endangering lives.

According to a study by the World Meteorological Organization (WMO) [Zhongming et al. 2021], 50% of all natural and technological disasters worldwide are related to weather, climate, and water hazards. These disasters resulted in the loss of 2.07 million lives and $3.6 trillion in the economy from 1970 to 2019. During this period, the highest economic losses were associated with storms and floods.

In South America, more than half of the disasters are related to floods, resulting in approximately $60 billion in economic losses on the continent between 1970 and 2019. The second-largest natural disaster in terms of lives lost occurred in Brazil in 2011, a flood that resulted in the loss of over 900 lives [Zhongming et al. 2021].

In the northern region of Brazil, particularly in the state of Pará, more than 900 hydrological disasters, such as heavy rainfall and floods, occurred between 1991 and 2021, according to data from the Digital Atlas of Deforestation in Brazil [Brasil 2022]. These

disasters caused approximately R$ 2.4 billion in damages and resulted in the loss of hundreds of lives.

These severe weather events can develop rapidly, posing challenges for forecasting and appropriate decision-making responses. There is a crucial need for the improvement of systems capable of predicting these events, providing early warnings and accurate information about their location, intensity, and duration, thus enabling effective response and enhanced mitigation measures to minimize the negative impacts.

Deep learning is an approach that has been used to enhance this response. When using weather radar data for precipitation nowcasting, models ranging from the use of convolutional layers in long short-term memory (LSTM) networks [Shi et al. 2015], the adaptation of U-Net for precipitation nowcasting [Ronneberger et al. 2015, Ayzel et al. 2020], to the utilization of attention mechanisms commonly employed in modern deep learning models have been highlighted [Vaswani et al. 2017, Trebing et al. 2021, Gao et al. 2022].

The U-Net architecture is commonly employed in precipitation nowcasting, when it is treated as an image-to-image translation problem, i.e., when a model maps a sequence of input images, at certain points in time, to a sequence of output images, at certain points in the future. In addition, the U-Net architecture has been shown to achieve better results in precipitation nowcasting than traditional numerical methods, when assuming a prediction window in the order of a few hours [Agrawal et al. 2019].

Therefore, this study aims to address three key questions. The first question (i) evaluates whether a U-Net architecture can effectively perform precipitation nowcasting, in terms of reflectivity (dBZ), for the next hour given the previous hour of meteorological events. The second question (ii) investigates different reflectivity thresholds and the results obtained by evaluation metrics. The third and final question (iii) aims to compare the observed and predicted data visually in relation to reflectivity thresholds and demonstrate how these visual results correlate with evaluation metrics.

This work is structured as follows: Section 2 presents the related works. Section 3 presents the materials and methods employed in this study, including the data and dataset building, architecture and training configuration, and precipitation nowcasting evaluation metrics. Section 4 presents and discusses the results. The conclusions drawn from the work are summarized in Section 5, along the discussion of future steps and potential research areas.

2. Related works

One of the pioneering works in solving the problem of precipitation nowcasting as a spatio-temporal sequence is the study by [Shi et al. 2015]. They extended the recurrent neural network (RNN) architecture known as LSTM by adding convolutional structures to the state transitions, resulting in the architecture called ConvLSTM.

Another relevant work is the Trajectory Gated Recurrent Unit (TrajGRU) model, presented by [Shi et al. 2017], which actively learns the location-variant structure for recurrent connections. It addresses the limitation of the previous work [Shi et al. 2015], which is location-invariant, while natural movements and transformations are generally location-variant. The study by [Kim et al. 2017] also utilizes ConvLSTM, proposing the

use of four depth channels to represent different elevations of data captured by weather radar.

Regarding the U-Net architecture, the RainNet model proposed by [Ayzel et al. 2020], is one of the first to adapt it for precipitation nowcasting. The U-Net is used here to predict the next prediction interval based on the previous four intervals, with each interval corresponding to 5 minutes. Originally applied to biomedical image segmentation [Ronneberger et al. 2015], the U-Net performs the task of image-to-image translation, similar to autoencoder, through the encoder and decoder components, in this way, U-Net outperforms traditional numerical methods in a window of a few hours [Agrawal et al. 2019].

The work by [Bonnet et al. 2020] employs the PredRNN++ model to predict future sequences of weather radar reflectivity images. PredRNN++ utilizes the Causal LSTM recurrent structure and the Gradient Highway Unit to address the gradient vanishing problem.

A recent study by [Ravuri et al. 2021] adopts an innovative approach to short-term precipitation nowcasting using conditional generative adversarial networks (cGANs). The results demonstrate that the model learns the distribution of weather radar data and can generate realistic future samples, capturing complex patterns and improving precipitation nowcasting accuracy.

Two notable works explore the use of attention mechanisms. The study by [Trebing et al. 2021] proposes the SmaAt-UNet architecture, which combines the U-Net structure with attention mechanisms for precipitation nowcasting. The work by [Gao et al. 2022] introduces Earthformer, which employs spatio-temporal transformers for predicting terrestrial systems, including precipitation. Both approaches are innovative and make significant contributions to precipitation nowcasting.

3. Materials and methods

3.1. Data

The data used in this study was collected from the X-Band weather radar of a mining company located in Canaã dos Carajás, Pará, Brazil. The radar covers a radius of 150 km, which corresponds to latitudes from -7.4251 to -4.7280 and longitudes from -51.4222 to -48.7250. Additionally, the radar generates 300×300 resolution scans/images every 5 minutes. With these characteristics, the weather radar is capable of generating 287 images in a single day of operation.

From the measurements obtained by the radar, several variables can be computed. However, in this study, we use the maximum reflectivity for short-term precipitation forecasting, referred to here as reflectivity. Reflectivity is obtained by taking the maximum reflectivity value from each of the scans at the ten elevations of the weather radar.

Figure 1 shows the coverage radius of the weather radar, encompassing municipalities such as Canaã dos Carajás, Parauapebas, Curionópolis, and Marabá. The coverage range is represented by the green rectangle on the map. The figure also includes two points representing two important iron ore mines, namely Serra Norte and S11D, represented by the purple and red markers, respectively.

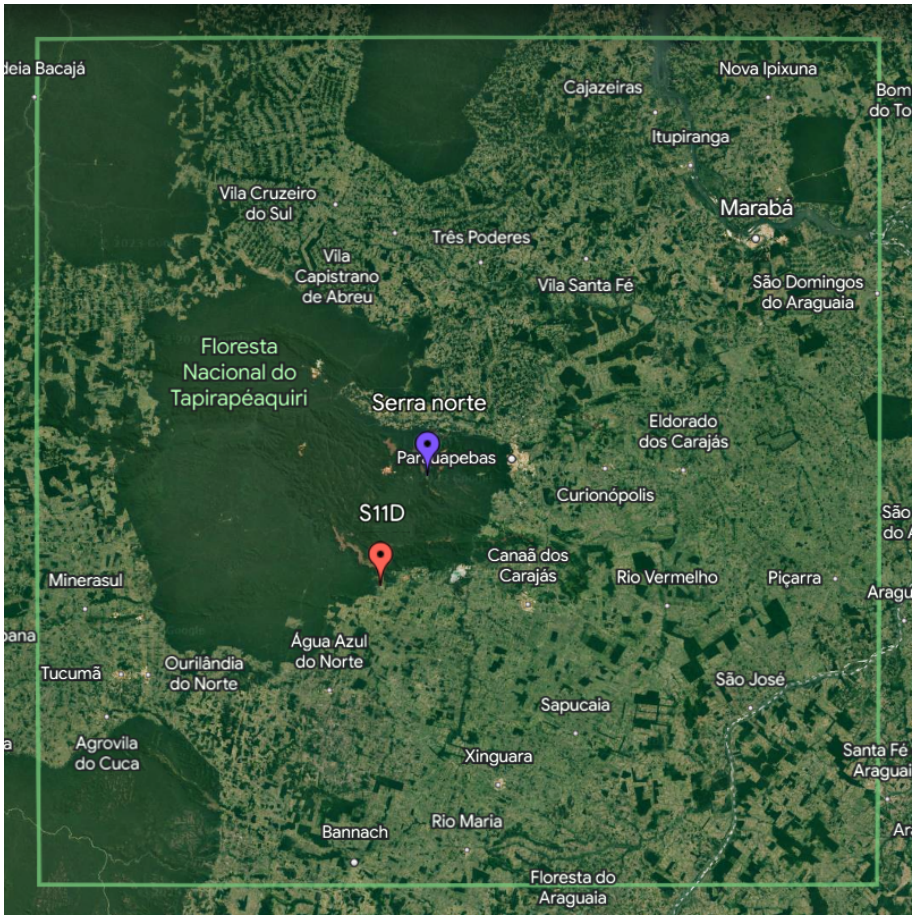


Figure 1. Weather radar coverage area. The purple and red markers show two important iron ore mines, which are Serra Norte and S11D, respectively.

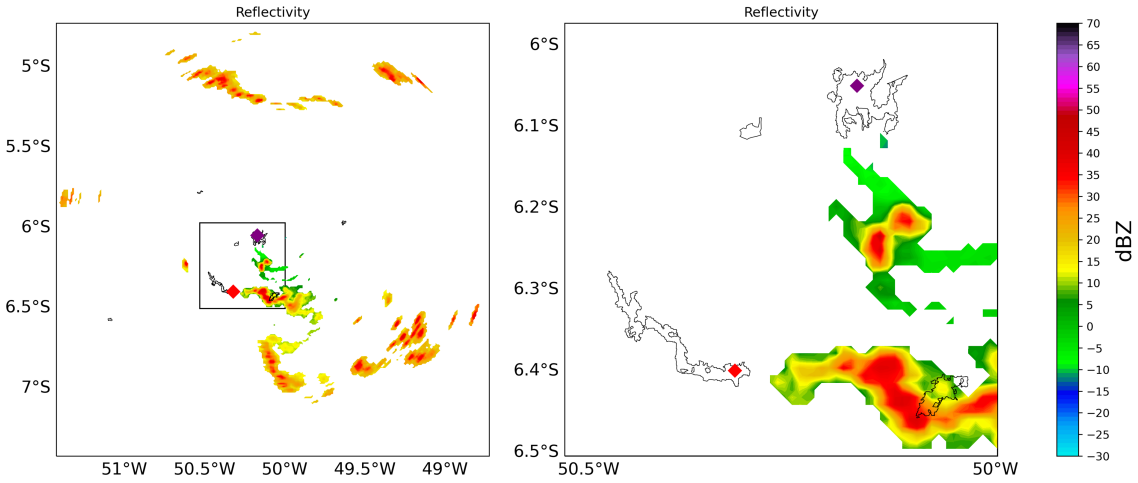
3.2. Dataset build

To further investigate the points shown in Figure 1, a region of interest encompassing both points was delimited. Figure 2 presents a meteorological event obtained by the radar at a specific time. Figure 2a displays the original radar data, with a resolution of 300×300, where the purple and red points represent the points of interest, and the black rectangle represents the region of interest. On the other hand, Figure 2b shows the region of interest, with a resolution of 60×60, and the points to be analyzed within it.

Three steps are performed to create the reflectivity image sequences that compose the dataset used in this study. The steps are sliding windows, sequence building, and removal of useless sequences. Figure 3 presents the schematic diagram of the three steps that compose the dataset build.

In the first step, window sliding, a window of size 60×60 is shifted with a step of 60 (width and height) throughout the original image. The resolution of the data after sliding window is $M \times N \times W \times W$, where M is the number of images, W is the window size, and N is the number of sub-images or windows in each image, given by Equation 1.

$$N = \left(\frac{H}{W}\right)^2 + 1 \tag{1}$$



(a) Original data (b) Region of interest

Figure 2. Perspectives of a meteorological event, measured in terms of reflectivity (dBZ), at a specific time. Original radar data with a region (rectangle) and points (purple and red markers) of interest (a), and region of interest (zoomed), 60×60 resolution, with the points of interest (b).

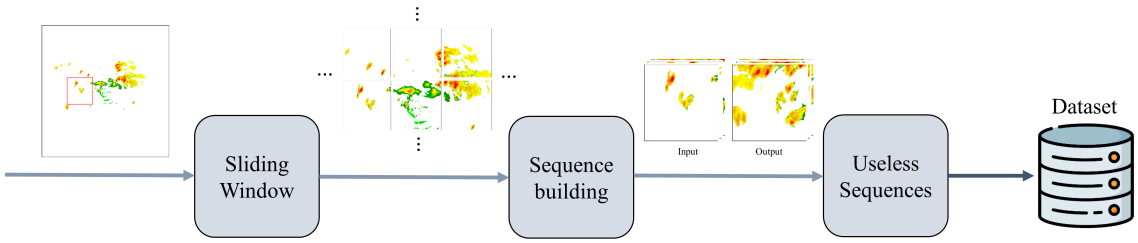


Figure 3. Schematic diagram showing the dataset build flow with each steps.

Where H is the height/width of the original image, and +1 represents the addition of the region of interest, which is not captured originally by the sliding. Therefore, for a complete day of 287 images with a resolution of 300×300, we obtain an output resolution of $287 \times 26 \times 60 \times 60$, meaning that for each image, we have 26 sub-images with a resolution of 60×60.

To perform precipitation nowcasting using weather radar data, it was defined that the prediction for the next hour is based on the previous hour. In other words, the model input consists of 13 images (previous hour), and the output consists of 12 images (next hour), where each image represents a 5-minute interval. Therefore, the input ranges from minute 0 to minute 60, while the output ranges from minute 65 to 120, both with a 5-minute interval.

Thus, it is possible to slide between the intervals and generate the image sequences that compose the model's inputs and outputs. The output resolution of the sequence building step (second step) can be defined as $S \times W \times W \times IN/OUT$, where IN and OUT are the numbers of input and output images, respectively, and S is the number of sequences given by Equation 2:

$$S = (M - (IN + OUT))\,(N + 1) \tag{2}$$

Hence, for a full day composed of 287 images, after the sliding window and sequence building steps, the data is obtained with the resolution of $7074 \times 60 \times 60 \times 13$ for the model's input and $7074 \times 60 \times 60 \times 12$ for the model's output.

The third and final step aims to remove useless sequences from the dataset formed so far. From the sliding window input of Figure 3, it can be observed that there are several areas in the image where no reflectivity measurements are obtained by the radar. Therefore, sequences formed by such images are excluded from the dataset as they do not contribute or negatively affect the precipitation nowcasting model.

Considering that the locations with reflectivity measurements are considered rare events, as they represent a much smaller portion of the image compared to the rest where no reflectivity measurements (Not a Number - NaN values) exist. Thus, after this step, a significant number of sequences are removed from the dataset, which depends on the day and the intensity of the meteorological event captured on that day.

3.3. U-Net architecture

To capture the behavior of meteorological events in the dataset, a U-Net architecture is used for precipitation nowcasting using weather radar data. Figure 4 presents the U-Net architecture used in this work. The network input is a sequence of 13 images, and the output is a sequence of 12 images, both with a resolution of 60 x 60.

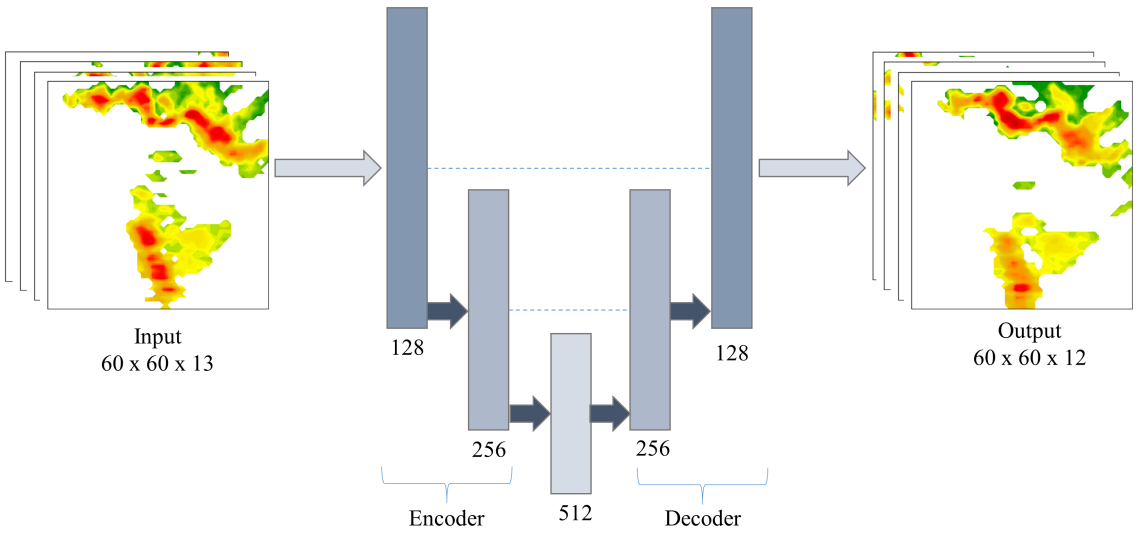


Figure 4. U-Net Architecture. The input to the architecture is a sequence of 13 images, which flow through the encoder, central, and decoder components, resulting in an output sequence of 12 images.

The U-Net architecture consists of two components: the encoder and the decoder. The encoder is composed of two blocks of operations, with each block representing a sequence of two repetitions of the following operations: convolution, batch normalization, and activation function. The number of filters in the convolutions is represented by the increasing numbers in each block (as shown in Figure 4), and a kernel size of 3 is used.

The ReLU (Rectified Linear Unit) activation function is employed. Each encoder block concludes with a max pooling operation with a pooling size and stride of 2, aiming to reduce the dimensionality while retaining important information.

The central block in Figure 4, with 512 filters, follows the same sequence of operations as the encoder, but without the max pooling at the end. Finally, the decoder is also composed of two blocks of operations, differing from the encoder by the absence of max pooling and the inclusion of two additional layers at the beginning: transposed convolution (or deconvolution) and concatenation. The transposed convolution layer aims to expand/increase the dimensionality with kernel size and stride of 2. The concatenation layer connects (as shown by the dashed line in Figure 4) each decoder block to its corresponding encoder block.

In the decoder, the sequence of two repetitions of operations in each block behaves in a reverse manner compared to the encoder. The number of filters increases through the blocks, while the other parameters remain the same. Finally, to generate the output image sequence, an output convolutional layer is inserted, consisting of 12 filters, a kernel size of 1, and a linear activation function.

3.4. Training setup

The model training was performed using the Adam optimizer and mean squared error (MSE) was used as the loss function. MSE was chosen as the cost because it measures the average squared difference between the predicted and observed reflectivity values.

To standardize the training and testing data of the model on a similar scale, z-score standardization was used. The z-score aims to standardize the data to resemble a normally distributed Gaussian distribution with a mean of zero and a standard deviation of one. Equation 3 presents the z-score:

$$z = \frac{x - \mu}{\sigma} \tag{3}$$

Where x is the unstandardized data, μ is the mean of x, σ is the standard deviation of x, and z is the standardized data. For model training, the mean and standard deviation used for standardization are obtained from the training data and applied to both training and testing data.

3.5. Forecast evaluation

In order to assess the quality of the forecast, given a certain criterion, and not the value of the forecast itself, metrics are used that categorize the investigated weather event. In this sense, the value of the forecast at a given point may be erroneous (large difference between observed and predicted values), but given a threshold value criterion, for example, it is possible to investigate the behavior of a storm in a region of interest, through information on the values that are above or below the chosen threshold. Finally, the analysis can be valuable in helping the forecaster to issue higher quality warnings.

The accuracy (ACC) and critical success index (CSI) are the two metrics used to measure the performance of precipitation forecasting. Both metrics require the categorization of the investigated meteorological events as either ”occur/yes/1” or ”do not

occur/no/0," for example. To achieve this, a reflectivity threshold needs to be defined to categorize the meteorological events that are above ("occur") or below ("do not occur") that threshold. In this context, an individual reflectivity value forming an image is considered a meteorological event.

Four combinations of prediction and observation categorizations are needed to calculate the ACC and CSI metrics: hit, miss, false alarm, and correct negative. A hit occurs when both the observed and predicted events are categorized as "occur." A miss is defined as the combination where the observed event "occurs," but the predicted event "does not occur." A false alarm occurs when the observed event "does not occur," but the prediction indicates that it "occurs". Lastly, a correct negative is when both the observed and predicted events are categorized as "do not occur".

Thus, ACC can be defined by Equation 4:

$$ACC = \frac{hits + \textit{correct negatives}}{total} \tag{4}$$

Where the total is the sum of all possible event combinations. ACC can be interpreted as the fraction of predictions that were overall correct, where a higher value indicates a higher percentage of correct predictions.

The CSI metric measures how well the predicted "occur" events correspond to the observed "occur" events and can be calculated using Equation 5.

$$CSI = \frac{hits}{hits + misses + \textit{false alarms}} \tag{5}$$

3.6. Experiments setup

Three full days of data were used to train the U-Net model for precipitation forecasting. After all the steps of dataset building, a total of 4820 input-output sequences, denoted as X and y, were generated. The input sequences X consists of 13 images, while the output sequence y consists of 12 images, both with a resolution of 60×60.

To ensure better data definition for the precipitation nowcasting model, all reflectivity values less than zero are considered zero. This includes NaN values (reflectivity absence).

The dataset is divided into 80% for training and 20% for testing, resulting in 3856 training input-output sequences and 964 testing input-output sequences. The mean and standard deviation are calculated based on the training input data and then applied to both the training and testing input-output data. The mean μ and standard deviation σ values for reflectivity are 4.06 and 7.96 dBZ, respectively.

In the evaluation of precipitation nowcasting, three different reflectivity thresholds are investigated, which are applied to the ACC and CSI metrics. These thresholds are 10, 15, and 20 dBZ. Consequently, the ACC and CSI results are calculated for each prediction interval, i.e., for each image/time that forms the sequence of 12 images.

Furthermore, the observed and predicted meteorological events are compared visually, as well as the categorical combination of observed and predicted events using each

reflectivity threshold.

4. Results and Discussion

Figure 5 presents the comparison between the observed and predicted images for four prediction intervals: 65, 85, 105, and 120 minutes, corresponding to 1st, 5th, 9th, and 12th prediction intervals, respectively.

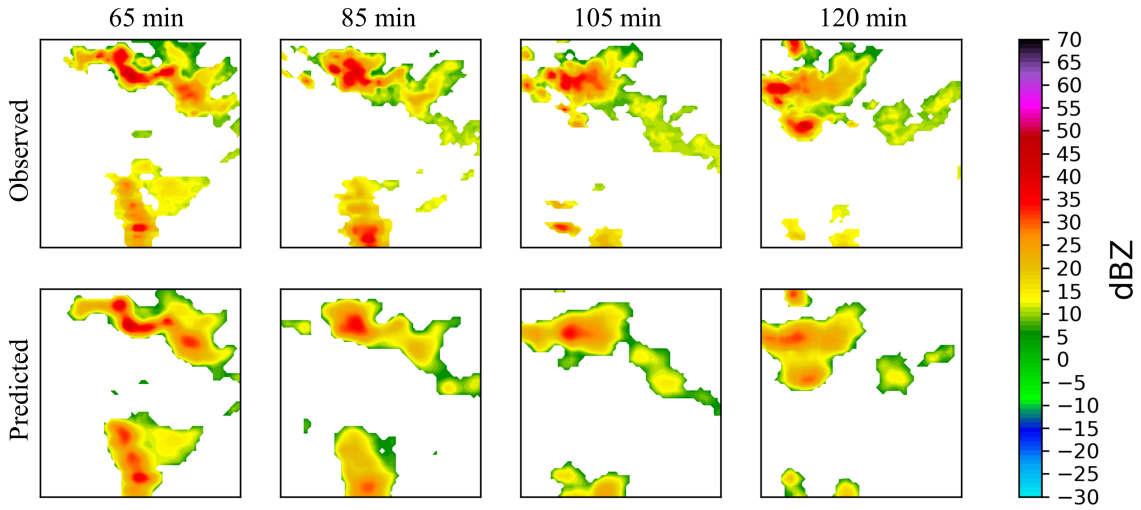


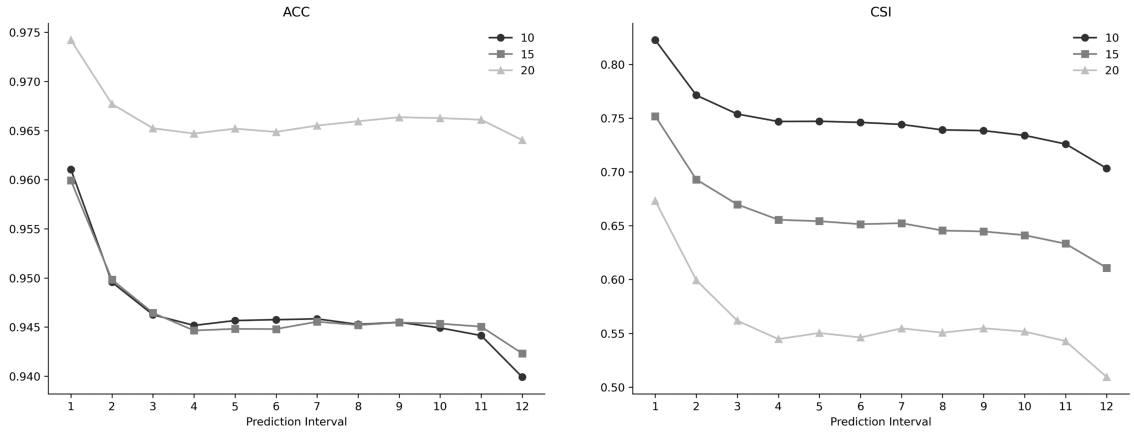
Figure 5. Observed and predicted images by the U-Net model. The precipitation nowcasting for the next hour is shown in the prediction intervals of 65, 85, 105, and 120 minutes.

It can be observed in Figure 5 that the prediction is capable of satisfactorily capturing the spatio-temporal behavior of the meteorological event in question. The model's prediction can also identify and simulate the trajectory, which is a crucial characteristic in precipitation nowcasting. However, when a pixel-to-pixel comparison is performed, it is noticeable that the U-Net does not accurately identify the intensity of reflectivity (rain). Finally, there is a decrease in prediction performance as the prediction interval increases, indicating that the further the prediction is from the input images/intervals, the lower the performance.

The line graphs in Figure 6 show each evaluation metric per prediction interval for each of the reflectivity thresholds. Figure 6a demonstrates that the best performance in terms of ACC is achieved with the 20 dBZ threshold, with an average (prediction intervals) of 0.9663, indicating that 96.63% of the predictions were correct. On the other hand, there is a similarity between the results for the 10 and 15 dBZ thresholds.

The behavior of the results in Figure 6a differs from Figure 6b, as the 10 dBZ threshold yields the best result in terms of CSI, indicating an average of 75.82% correct "occur" events. Additionally, the 20 dBZ reflectivity threshold performs the worst among the three thresholds, achieving only an average of 56.22%.

The results of both metrics (Figure 6) for the 20 dBZ reflectivity threshold suggest that the ACC result can be misleading, as it is sensitive to correct negatives, while CSI is only sensitive to hits. Therefore, it can be inferred that there is not a large number of



(a) Accuracy (b) Critical Success Index

Figure 6. Evaluation metrics for precipitation nowcasting in prediction intervals for each of the reflectivity thresholds used, 10, 15, and 20 dBZ.

reflectivity values above 20 dBZ, thus increasing the number of correct negatives (values lower than 20 dBZ or NaN).

Figure 7 is presented to support the inference made from the results in Figure 6. It shows the comparison of results for one prediction interval in terms of the categorical combinations of observed and predicted events, which generate masks for each reflectivity threshold.

It can be noted that the mask in Figure 7c (10 dBZ) shows the highest similarity between the categorical combinations of observed and predicted events. The contour generated by the hits satisfactorily represents the event in question. When analyzing the 20 dBZ threshold (Figure 7e), it can be observed that the mask successfully captures the events (observed and predicted) above the threshold, although with a lower number of hits compared to Figure 7c.

The decrease in the number of hits and the increase in correct negatives shown in Figure 7e support the evidence that, in general, there are many reflectivity values below 20 dBZ in the dataset, including NaN values. Thus, ACC may not be a reliable metric, as it is highly sensitive to the category of the most common events, the correct negatives.

5. Conclusions and future works

This study aims to evaluate the capability of a U-Net model in precipitation nowcasting using weather radar data. Additionally, three types of reflectivity thresholds are investigated, along with the results obtained by these thresholds and the visual correlation between observed, and predicted data and the evaluation metrics for precipitation nowcasting.

Analyzing the results, it can be observed that the prediction results are satisfactory, as the model efficiently captures the trajectory and intensity of the investigated weather events. When observing the reflectivity thresholds, it is noticed that the ACC result obtained by the 20 dBZ threshold can be misleading, as it is sensitive to correct negative events (values lower than 20 and NaN), which are the most common events and are supported by the masks generated by the observed and predicted data for each investigated

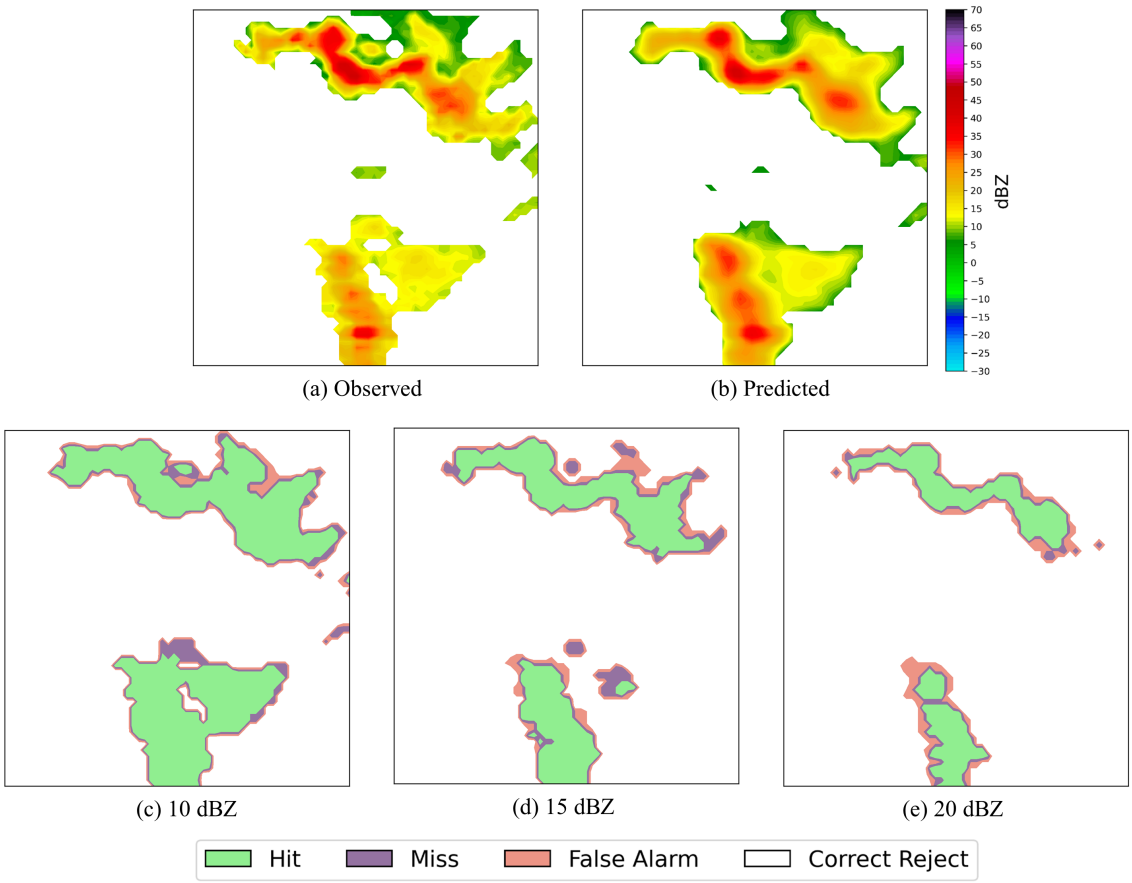


Figure 7. Analysis of the categorical combinations of observed (a) and predicted (b) events in terms of the masks generated by them using the reflectivity thresholds of 10 (c), 15 (d), and 20 (e).

threshold.

These results indicate the absence of a large number of reflectivity values greater than 20 dBZ in the built dataset, suggesting that the 10 and 15 dBZ thresholds may be more suitable for analyzing the precipitation nowcasting results. The result obtained with a 10 dBZ threshold stands out, reaching a CSI of 0.7582, indicating that more than 75% of values above the threshold were correctly predicted.

For future work, a more detailed statistical investigation can be conducted and incorporated into the dataset building steps to obtain an ideal reflectivity threshold value. Additionally, other deep learning architectures can be tested to enhance the precipitation nowcasting.

References

Agrawal, S., Barrington, L., Bromberg, C., Burge, J., Gazen, C., and Hickey, J. (2019). Machine learning for precipitation nowcasting from radar images. *arXiv preprint arXiv:1912.12132*.

Ayzel, G., Scheffer, T., and Heistermann, M. (2020). Rainnet v1. 0: a convolutional neural network for radar-based precipitation nowcasting. *Geoscientific Model Development*, 13(6):2631–2644.

Bonnet, S. M., Evsukoff, A., and Morales Rodriguez, C. A. (2020). Precipitation nowcasting with weather radar images and deep learning in são paulo, brasil. *Atmosphere*, 11(11):1157.

Brasil (2022). Ministério do desenvolvimento regional. Secretaria de Proteção e Defesa Civil. Universidade Federal de Santa Catarina. Centro de Estudos e Pesquisas em Engenharia e Defesa Civil. Atlas Digital de Desastres no Brasil. Brasília: MDR, 2022.

Gao, Z., Shi, X., Wang, H., Zhu, Y., Wang, Y. B., Li, M., and Yeung, D.-Y. (2022). Earthformer: Exploring space-time transformers for earth system forecasting. *Advances in Neural Information Processing Systems*, 35:25390–25403.

Kim, S., Hong, S., Joh, M., and Song, S.-k. (2017). Deeprain: Convlstm network for precipitation prediction using multichannel radar data. *arXiv preprint arXiv:1711.02316*.

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., et al. (2021). Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer.

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

Shi, X., Gao, Z., Lausen, L., Wang, H., Yeung, D.-Y., Wong, W.-k., and Woo, W.-c. (2017). Deep learning for precipitation nowcasting: A benchmark and a new model. *Advances in neural information processing systems*, 30.

Trebing, K., Stanczyk, T., and Mehrkanoon, S. (2021). Smaat-unet: Precipitation nowcasting using a small attention-unet architecture. *Pattern Recognition Letters*, 145:178–186.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Zhongming, Z., Linong, L., Xiaona, Y., Wangqiang, Z., Wei, L., et al. (2021). Atlas of mortality and economic losses from weather, climate and water extremes (1970-2019). *WMO*.