Precipitation Nowcasting using a Generative Adversarial Network

S. van Os

Delft University of Technology

TU Delft HKV

Precipitation Nowcasting using a Generative Adversarial Network

by

S. (Sven) van Os

In partial fulfilment of the requirements for the degree of
Master of Science
in Civil Engineering
at Delft University of Technology
Track: Geoscience and Remote Sensing
to be defended publicly on June 13, 2024 at 11:00 AM (CEST)

Student number:	4448642	
Project Duration:	May, 2023 - June, 2024	
Thesis committee:	Dr. M.A (Marc) Schleiss,	TU Delft, dept. Geoscience & Remote Sensing (CiTG), Chair
	Dr. R. (Riccardo) Taormina,	TU Delft, dept. Water Management (CiTG)
	Prof.dr.ir. R. (Remko) Uijlenhoet,	TU Delft, dept. Water Management (CiTG)
	Dr. M. (Mattijn) van Hoek,	HKV Lijn in Water B.V.
	Ir. D. (Dorien) Lugt,	HKV Lijn in Water B.V.

Cover Image: Thunderstorm in Southwest, IA (reddit, user/el_dpalablo)

An electronic version of this thesis is available at http://repository.tudelft.nl/.

TU Delft

Preface

This thesis was written in partial fulfilment of the requirements for the degree of Master of Science in Civil Engineering at Delft University of Technology. Although this project took longer than expected and I had to reduce its scope due to unforeseen problems, I look back on this period with pride.

The completion of this thesis would not have been possible without the support of the following people. First of all, I would like to thank the members of my graduation committee. Marc Schleiss, thank you for introducing me to this topic and for your seemingly endless patience and support throughout the past year. Riccardo Taormina, your contagious enthusiasm greatly motivated me. Remko Uijlenhoet, your expert knowledge has been invaluable, and your reminders to step back and keep an overview were much appreciated. Mattijn van Hoek and Dorien Lugt, thank you for giving me the opportunity to undertake this project at HKV Lijn in Water B.V., as well as for your feedback, in-depth discussions, guidance, and for keeping an eye on me over the past year.

Furthermore, I would like to thank Thomas Stolp for teaching me your structured approach to setting up the model code and helping me get it to work on the GPU, something I had no prior experience with. Yanghuan Zou, thank you for sharing your experience with training Deep Learning models on DelftBlue with me. Daniel Blázquez Martín, thank you for providing me with the S-PROG forecasts and for the enjoyable discussions and exchange of ideas; I hope you learned as much from them as I did. Jerom Aerts, thank you for your knowledge on nowcasting and for the idea of how to adapt PAT to fit into the goals of this project.

Finally, I would like to thank my friends and family for providing a listening ear when I was facing problems or when I had a breakthrough to share enthusiastically. Tim, your enthusiasm and knowledge on a lot of weather-related topics greatly influenced my work. Pim and Menno, thank you for helping me stay motivated and for keeping an eye on me. Pepijn, thank you for the many study sessions where we could share our thesis problems with each other. Lastly, to my parents, who have always been there for me, thank you for your unwavering support.

S. (Sven) van Os
Delft, June 2024

Abstract

Nowcasting high-intensity precipitation is crucial for emergency services and municipalities when making weather-dependent decisions. This research implements and trains a deep generative model for nowcasting using a cleaned precipitation radar composite dataset spanning 15 years, with a 5-minute temporal and 1 km spatial resolution.

We propose and apply a method for speckle-like clutter removal to enhance data quality, particularly for high-intensity precipitation rates. The deep generative model is trained with two data sampling strategies to balance the data and improve the accuracy of high-intensity precipitation forecasts. Model performance is evaluated using several standard metrics, and we propose an adaptation to one metric to quantify a score to peak anticipation time in precipitation nowcasting. We also compare our model's performance with a state-of-the-art deterministic Lagrangian extrapolation-based nowcasting system.

Our results show that the proposed data quality improvement method effectively removes certain errors from historical radar data. Although the deep generative model currently scores low on the standard metrics, the model trained with a focus on high-intensity precipitation shows an improved score to peak anticipation time. Both deep generative models exhibit less blurring compared to the state-of-the-art model and, in some cases, perform similarly or outperform it.

Acronyms

AENN Adversarial Extrapolation Neural Network

CAPPI Constant Altitude Plan Position Indicator

cGAN conditional Generative Adversarial Network

CSI Critical Success Index

DGMR Deep Generative Model of Radar

ECMWF European Centre for Medium Range Weather Forecasts

ETH Echo Top Height

FAR False Alarm Rate

FSS Fraction Skill Score

GAN Generative Adversarial Network

ISW Importance Sampling Weight

KNMI Royal Netherlands Meteorological Institute

MAE Mean Absolute Error

MSE Mean Squared Error

NWP Numerical Weather Prediction

PAT Peak Anticipation Time

POD Probability of Detection

Contents

1

Introduction

High-intensity precipitation is a driver for many natural hazardous events that can cause severe damage to crops and infrastructure, disrupting society and can even cause loss of life (Ayzel et al., 2019; Douris et al., 2023; Manola et al., 2020; Xu et al., 2022). These high intensities are expected to increase in number in Western Europe due to a warmer atmosphere being able to hold more moisture (Lenderink and Van Meijgaard, 2010), which leads to the need for better early warning systems to support authorities in weather-dependent decision-making (Ji et al., 2023; Lin, 2022). Traditionally Numerical Weather Prediction (NWP) is used to forecast future precipitation based on governing equations of atmospheric dynamics and continuous data assimilation (Han et al., 2023). NWPs do however have limitations for short lead times (<6 h) in terms of accuracy, spatiotemporal resolution, and computation time for operational purposes (Berenguer et al., 2012; Imhoff et al., 2020; Pierce et al., 2012). Nowcasting is used to resolve the precipitation forecast for short lead times (Zhang et al., 2023). Modern precipitation nowcasting algorithms often rely on the extrapolation of observations by ground-based radars (Lebedev et al., 2019).

1.1. Research motivation

Nowcasting is traditionally done by estimating the apparent movement of radar precipitation fields using optical flow or variational echo tracking by extrapolating the observations into the future (Foresti et al., 2016; Grecu and Krajewski, 2000; Pulkkinen et al., 2019). The emergence of deep learning methods in the field of nowcasting provides new opportunities for developing new models that may learn to predict spatial and temporal structures of precipitation fields (Bi, 2022). Nowcasting deep learning methods have shown accurate prediction for low-intensity precipitation. However, they often produce poor results for high intensities and their predictions become blurry and unrealistic at longer lead times (Ravuri et al., 2021).

In many deep learning methods the extremes are often handled as outliers and are often ignored (Bi, 2022). This leads to the outputs from precipitation estimators being highly skewed towards lower values, resulting in fewer high-intensity precipitation in their predictions (Hayatbini et al., 2019). These extremes are however important to predict, as the

high-intensity precipitation can lead to flooding and loss of life. This leads to the need of the development of new algorithms that do not under represent these extremes.

To overcome the blurry predictions, the use of generative models was proposed to generate more realistic video prediction, which can be applied to radar echo extrapolation prediction (Liu and Lee, 2020; Wang et al., 2021; Xu et al., 2022). These generative models utilize a training strategy developed by Goodfellow et al. (2014) and this type of architecture is called a Generative Adversarial Network (GAN). These methods have shown great potential, and a deeper understanding of these models is required to improve their predictive capabilities for precipitation.

The main objective of this thesis is to enhance the understanding of deep learning nowcasting models regarding their predictive capability of high-intensity precipitation. Additionally, this research seeks to gain a deeper understanding of training data cleansing as well as the influence of training data sampling on these models. By achieving these aims, valuable insights will be provided, leading to improved algorithm development in the field of precipitation nowcasting.

1.2. Research objective

The aim of this research is to implement and train the Deep Generative Model of Radar (DGMR), as first described by Ravuri et al. (2021), on a new radar dataset cleaned of speckle-like clutter as well as gain more insight into the effect of data sampling during training of deep learning nowcasting models. This is achieved by analysing and processing the training data and by applying two training strategies with the aim of improving the capability of DGMR to predict high-intensity precipitation events in the Netherlands. The effectiveness of the training strategies is evaluated with metrics and compared to S-PROG as benchmark on some events to show some of the strengths and limitations of these training strategies. The code for the model was taken from Elsmann (2023).

1. **How can the data be processed to improve the data quality for high-intensity precipitation events in the Netherlands?**
2. **In what ways can a Generative Adversarial Network (GAN) be trained with radar images to improve the prediction for high-intensity precipitation events in the Netherlands?**
3. **How does this model, under different training strategies, compare to S-PROG, a state-of-the-art extrapolation based nowcasting system?**

2

Related research

In Section 2.1 the concept of Generative Adversarial Network (GAN) is introduced and some earlier work in weather prediction using GANs are highlighted. The models used in this research are introduced in Section 2.2.

2.1. Generative Adversarial Networks background

A commonly used training strategy of machine learning is supervised learning, where a dataset of example inputs and example outputs are used to learn to map the input to an output (Goodfellow et al., 2020). Generative models do not follow this training method as they use an unsupervised learning approach. They have the goal to study a collection of training examples and learn the probability distribution that is used to generate these examples to make predictions. The training of these networks is however difficult and for this reason adversarial networks have been proposed (Goodfellow et al., 2014). Here a generative model and an adversary, a discriminative model, compete against each other in a minimax game, which is a Generative Adversarial Network (GAN).

In simple terms, the generator's objective is to generate samples from noise that can fool the discriminator (Schreurs, 2021). This is done with a mapping function, $G(z;\theta^{(G)})$, that takes noise, z, with a set of learnable parameters, defined by $\theta^{(G)}$, so it can map noise to realistic samples (Goodfellow et al., 2020) and it tries to minimize how often the discriminator correctly labels the generated samples.

The discriminator's objective is to detect if a presented sample is a generated sample or a real sample (Schreurs, 2021). The discriminator does this by examining real and generated samples, x, and returns an estimate $D(x;\theta^{(D)})$. The discriminator tries to maximize correctly labeling the presented samples (Goodfellow et al., 2020).

In this adversarial setup, the generator, G, and discriminator, D, are put into a two-player minimax game, where the aim is to optimize the value function, $V(G,D)$, representing the skill of each model. This value function is given by:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[logD(x)] + \mathbb{E}_{x\sim p_z(z)}[log(1-D(G(z)))] \tag{2.1}$$

Variations of this training approach have been used to improve generative models for different tasks, including video prediction (Clark et al., 2019) and text-to-image generation (Zhang et al., 2016). Several methods for the training process of GANs with a variety of applications have been proposed to stabilize the training process (Arjovsky et al., 2017; Brock et al., 2018; Karras et al., 2017; Saito et al., 2018; Salimans et al., 2016).

2.1.1. Earlier GAN applications in weather prediction

Bihlo (2020) uses a conditional Generative Adversarial Network (cGAN) to learn the physics underlying for the geopotential height of the 500 hPa pressure layer, the two-meter temperature and the total precipitation. The cGAN is fed input data to condition the output on by giving more context to the model of the initial conditions it uses to make a forecast (Mirza and Osindero, 2014). In the case of Bihlo (2020) this conditional input is European Centre for Medium Range Weather Forecasts (ECMWF) data of the past day to predict the next 24 hours for the given input feature. The generator follows a U-Net architecture based on Isola et al. (2018) and initially proposed by Ronneberger et al. (2015). The same model architecture is retrained for each feature by training on that feature alone. To investigate whether meaningful statistical information could be obtained, they run an ensemble of slightly different cGANs by introducing Monte-Carlo dropout layers into the model architecture. Three case studies show that the forecasts made for the geopotential height and two-meter temperature captures the ground truth well for all lead times. However the total precipitation is not always correctly captured in the ensemble for all lead times. This indicates that the total precipitation is difficult to forecast correctly for these lead times using only the total precipitation as context for the model. These results have not been compared to another model as a benchmark.

Choi and Kim (2022) designed a cGAN based model for advanced precipitation nowcasting with good prediction performance for dam basins in South-Korea. The generator is based on the same U-net architecture from Isola et al. (2018) and is trained for each dam basin separately, using different transfer learning strategies to develop a precipitation nowcasting model for different dam basins. The generator only predicts the next frame and by applying a recursive process it achieved adequate performance up to 80 minutes, where other shown state-of-the-art models could only achieve this performance up to 60 minutes. However, their model has the tendency to underestimate intense precipitation events, which may be due to data imbalance as there are few intense precipitation events in the training set compared to lower precipitation rates.

Duncan et al. (2022) applied a GAN that integrates multi-scale semantic structure and style information to allow them to synthesize physically realistic fine-scale precipitation features with realistic high-intensity precipitation. This is done by fusing features at different spatial scales in the generator and have separate discriminators for each spatial scale to improve their predictive capability. The model outperforms a leading NWP model in skill up to 1-2 day lead time. Their analysis shows that their models as well as the used NWP underestimate the extremes that occur in the ground truth.

Jing et al. (2019) developed Adversarial Extrapolation Neural Network (AENN), which uses the last 5 precipitation radar frames to return an extrapolation for 30, 60 and 90 minutes. They have shown that their model outperforms other models significantly and it can generate accurate and realistic extrapolation echoes by making use of GAN to avoid blurry predictions. The model was trained and tested on five CINRAD/SA doppler weather radars provided by the National Meteorological Information Center of China. The examples show

that the data contains several radar error sources which may have influenced the model performance as well as the evaluation. Furthermore, the quantitative evaluation does not compare the performance of AENN to the other benchmark models for high precipitation rates. This model has also been applied to weather radar in the Netherlands by Schreurs (2021), where the inclusion of the adversarial loss during training has been demonstrated to reduce the blurriness compared to a training setup without the adversarial loss.

2.2. Nowcasting models

This research is performed using the GAN based deep learning model proposed by Ravuri et al. (2021) called Deep Generative Model of Radar (DGMR). The model performance is compared to the optical flow based extrapolation model proposed by Seed (2003) and implemented using PySTEPS (Pulkkinen et al., 2019). Both models and some earlier work with these models are described here.

2.2.1. Deep Generative Model of Radar (DGMR)

Ravuri et al. (2021) developed DGMR, a conditional generative model for precipitation nowcasting. Their model takes four consecutive radar observations, the past 20 minutes, and use these as context to generate multiple forecasts for the next 18 frames, 90 minutes. The generator of the model takes the context frames and latent variables as input and is trained using two discriminators and a regularization term. A spatial discriminator is used, which is a convolutional neural network that tries to distinguish individual observed radar frames from generated frames to ensure spatial consistency and discourage blurry predictions. A temporal discriminator is used, which is a 3D convolutional neural network which tries to distinguish observed and generated radar sequences to impose temporal consistency. The regularization term is used to further improve accuracy by penalizing deviations at a grid cell level between the radar sequences and the model predictive mean as computed with multiple samples.

The model was trained and evaluated on a radar composite over the United Kingdom from the Met Office RadarNet4 network with 15 C-band dual polarization radars (Ravuri et al., 2021). A schematic overview of the model training can be seen in Figure 2.1. After training, only the generator is used to make forecasts. The model was evaluated against PySTEPS, a radar only version of the MetNet model from Sønderby et al. (2020) and the U-Net encoder-decoder model similar to Agrawal et al. (2019) and Ayzel et al. (2020). Both a quantitative verification with commonly-used verification measurements was performed as well as a qualitative assessment with expert forecasters. However the U-Net model was not used in the qualitative assessment with expert forecasters. According to Ravuri et al. (2021) their generative model was judged to be more accurate and useful than PySTEPS or the MetNet model according to the professional forecasters. According to the quantitative evaluation, the DGMR is competitive compared to the baseline models, provides more accurate probabilistic forecasts and preserves the statistical properties of precipitation across spatial and temporal scales without blurring.

Frenkiel (2022) validated the generator of the pre-trained DGMR on the weather radar data from 2021 in the Netherlands and compared it to forecasts made with S-PROG. They found that the model performance is in line with the performance claimed by Ravuri et al. (2021), but that improvements could be made when the model is re-trained on the Dutch radar data. Elsmann (2023) trained DGMR on rain gauge adjusted radar data from 2008 up to

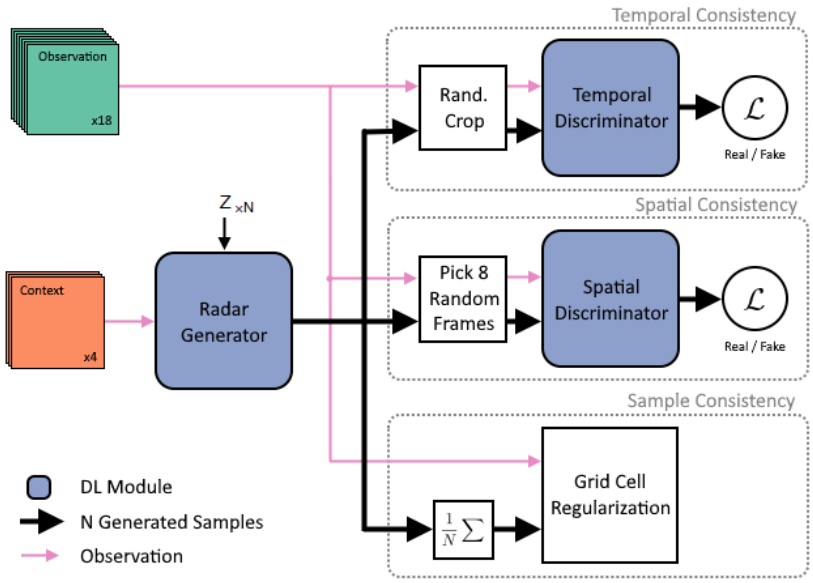

Figure 2.1.: Schematic diagram of the model architecture showing the generator with spatial latent variables Z. Image taken from Ravuri et al. (2021).

and including 2015 in the Netherlands. They also made a version of the model which takes Echo Top Height (ETH) as extra input for context. More information on this dataset can be found in Appendix E. They claim that the dataset is made up of a radar composite using 2 C-band radars, however one of the specified locations only became operational in 2016. The two DGMR versions were quantitatively evaluated on a set of verification metrics, however without comparing them to the performance of a benchmark model. The version including ETH as extra input showed an improved forecast for low precipitation rate events as well as large scale high precipitation rate events. Forecasts for small-scale high precipitation rates did not improve.

2.2.2. Spectral prognosis (S-PROG)

The S-PROG model consists of three components, the estimation of the advection field, decomposition of the field into Fourier components and a model for the scale-dependent Lagrangian evolution of the field (Seed, 2003). The decomposition is done using fast Fourier transformation and transformed back to the spatial domain, resulting in a cascade of a number of levels representing a different scale. Separate second-order auto-regressive processes are applied to each cascade level to account for the dynamic scaling of precipitation (Pulkkinen et al., 2019). The Lagrangian evolution is applied to each cascade level and then summed together to get the forecast. Both the precipitation intensity and motion field are assumed to be stationary with Lagrangian persistence and can therefore be implemented without training the model.

The open-source implementation of S-PROG is done using PySTEPS from Pulkkinen et al. (2019) as a benchmark in this study.

3

Data

In Section 3.1 the radar reflectivity dataset is introduced. Some of the preprocessing done by the KNMI of the radar reflectivity dataset is shown in Section 3.2.

3.1. KNMI radar reflectivity data

The KNMI currently uses two dual-polarized C-band radar systems to measure the precipitation reflectivity. These radars make 14 scans every 5 minutes under elevation angles of 0.3, 0.4, 0.8, 1.1, 2.0, 3.0, 4.5, 6.0, 8.0, 10.0, 12.0, 15.0, 20.0, and 25.0 degrees. The reflectivity product used here is the archive of the real time radar reflectivity composites, named radar_tar_refl_composites, which only use the scans at 0.3, 1.1, 2.0 and 3.0 degrees. These scans are used to generate a single image for each radar at an equivalent elevation of 1500 m, the Constant Altitude Plan Position Indicator (CAPPI). The two radar images are then combined by applying a weighted average of the radar reflectivities (Wessels, 2006). The resolution of the product is 1×1 km and is provided every 5 minutes. This product is downloaded from the data platform of the KNMI and starts on January 1, 2008, at 00:00 UTC. The data up to December 31, 2022 at 23:55 UTC is used. The archive of the radar reflectivity composites is used as it contains the same data available to an operational nowcasting system.

There are several sources of error when measuring the precipitation reflectivity, an overview of which can be seen in Figure 3.1. Some of these error sources can give high reflectivity values in isolated pixels without precipitation and when the radome is wet, it can give strong attenuation and reduce the reflectivity values (Holleman and Beekhuis, 2005). These are some of the limitations of using radar reflectivity to determine precipitation intensities.

Two single-polarized C-band radars were initially located in De Bilt and Den Helder and these systems have been upgraded in 2016/2017 to the current radar systems. Due to high-rise buildings in the vicinity of the radar system in De Bilt, the radar system was relocated to Herwijnen to reduce the ground clutter and shielding that these buildings caused. Both radar systems received several upgrades during this period to enhance the quality by using polarimetric weather radars and replacing the radar sensors with modern low-maintenance

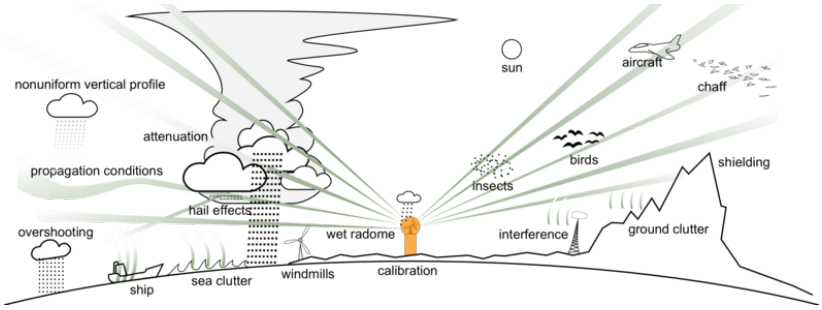

Figure 3.1.: Sources of error affecting radar measurement of precipitation from www.knmi.nl/research/observations-data-technology/projects/quality-enhancement-of-quantitative-precipitation-estimates.

radars (Leijnse et al., 2016). The extent of the radar reflectivity product with context of the old and current radar locations is illustrated in Figure 3.2.

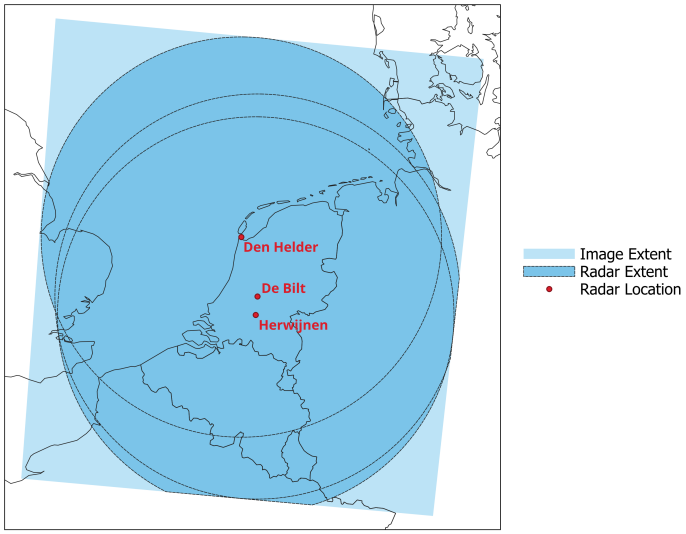

Figure 3.2.: Image extent (700×765 pixels) from the radar reflectivity product from the Royal Netherlands Meteorological Institute (KNMI) and the radar extents (320 km) for the old locations (De Bilt and Den Helder) and the current locations (Herwijnen and Den Helder).

The KNMI products are provided in HDF5 files, containing an 8-bit ($0 - 255$) 2D-array (700×765 pixels) with 0.5 dBZ intervals where the highest value of 255 represents a no-data pixel. These array values are transformed from their 8-bit values to dBZ with the transformation Equation 3.1.

$$dBZ = (\text{8-bit value} \times 0.5) - 32.0 \tag{3.1}$$

The Marshall-Palmer $Z - R$ relation, Equation 3.2, is often used as an approximation for the relation between the radar reflectivity factor Z and R in mm/h (Marshall and Palmer, 1948). Since the KNMI provides the data in dBZ, which is $10log_{10}(Z)$, the Marshall-Palmer $Z - R$

relation expressed in dBZ is used given by Equation 3.3 as defined by Wessels (2006).

$$Z = 200R^{1.6} \tag{3.2}$$

$$dBZ = 16 \log_{10} R + 23 \tag{3.3}$$

3.2. Data preprocessing

Holleman and Beekhuis (2005) describe how the fluctuations of the received power echoes in dB are analysed within each processed range bin. The standard deviation spectra of clutter and precipitation are different but overlap and this is used to filter out about 45% of the clutter signals. The clutter flags are set per range sample, of 1 by 1 km, depending on the observed standard deviation spectra and indicate that there may be clutter in that range sample. The two neighbouring range bins are also checked and the decision is made if the range sample contains clutter or not. A schematic overview of the clutter flag evaluation for a single sample can be seen in Figure 3.3 where it is shown represented in a conical view of the radar, with the view angle as the azimuth and the distance from the radar as the range. This method has been applied to the radar reflectivity dataset used in this research prior to the upgrade of the radars to Doppler systems, however it may fail to correctly flag clutter when it occurs within a precipitation field.

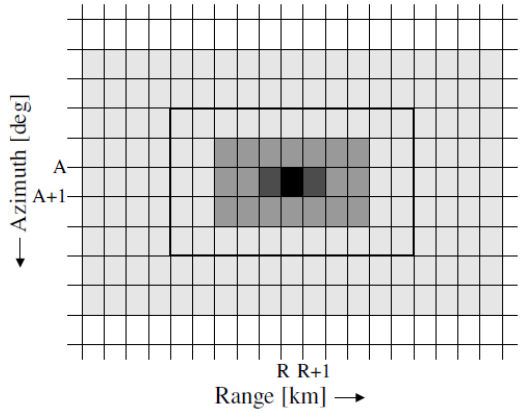

Figure 3.3.: Schematic view of the procedure on clutter removal. The central range bin and the two neighbours are marked in black and dark grey, respectively and are used for the clutter flag. This figure is taken from Leijnse et al. (2016).

Leijnse et al. (2016) describes a method for using the Doppler spectrum as a method of detecting clutter as it contains information on the distribution of radial velocities of the samples. Ground clutter can easily be recognized in a Doppler spectrum as it produces a narrow peak centered around the zero velocity, an example of which can be seen in Figure 3.4. The ground clutter is then removed by applying a steep high-pass filter to the Doppler signals frequency domain. Finally, a speckle filter is applied, where speckle indicates isolated range samples with valid data and neighbours with no data. This method was applied to the dataset since the radar systems have been upgraded on the Den Helder location and the introduction of the radar in Herwijnen.

As stated on the data platform of the KNMI, the volume scans of each radar generate a single image per radar that represents the reflectivity at 1500 meters. The two radar images are

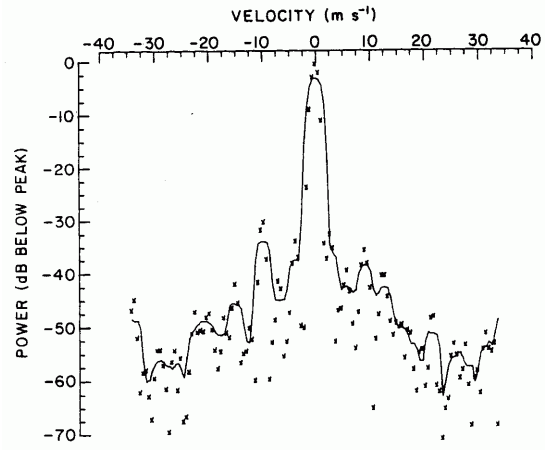

Figure 3.4.: A Doppler spectrum containing a peak indicating ground clutter centered at zero velocity. This figure is taken from Doviak and Zrnić (1993).

then combined into a single image by taking a weighted average of the radar reflectivities where the weights are a function of the distance to the radar. After the processing for the clutter removal and constructing the composite, the radar composite is made available on the data platform of the KNMI under the name radar_reflectivity_composites version 2.0.

An example of clutter that remains in the composite when applying the method described by Holleman and Beekhuis (2005) can be seen in Figure 3.5. A circular pattern with speckle-like clutter can be identified around the Den Helder radar station. However, the reflections measured to the East around the border with Germany move sporadically over time and are also caused by non-precipitation targets. Since nowcasting ideally is done on precipitation only, the non-precipitation targets will from now on be called errors. It can also be seen that some pixels with high precipitation rates within the larger precipitation field to the West do not move with the precipitation field itself. These areas representing high-intensity precipitation rates remain on the same location, but their intensity changes over time depending on the precipitation field that moves over those locations. This is most likely caused by ground clutter which has not been removed in this example as the method described by Leijnse et al. (2016) was not applied. The ground clutter is most pronounced over sea and influences the precipitation rates that are inferred.

The KNMI is continuously working on improving the clutter filtering and improving the quantitative precipitation estimation, further improving the quality of the radar product for precipitation nowcasting. See for a recent overview Overeem et al. (2020, 2021).

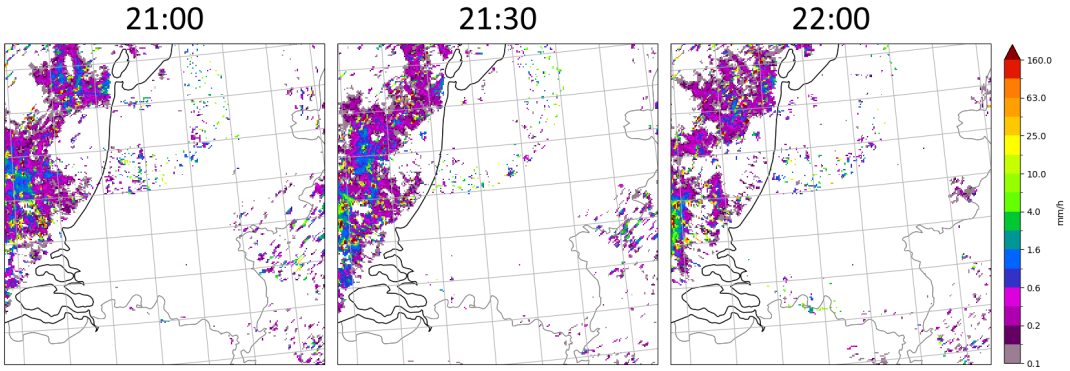

Figure 3.5.: Precipitation rates with speckle-like clutter in a circle around the Den Helder radar location as well as ground clutter representing more than 160 mm/h within the precipitation field over the North Sea based on radar reflectivity on October 31st, 2014, times given in UTC.

4

Methodology

First a method for removing speckle-like clutter in the historical radar data is introduced in Section 4.1. Then an event weighting method to use for selecting training data as well as how to use it during training is introduced in Section 4.2. The data preparation and training setup of DGMR are described in Section 4.3. Finally, the the verification methods used in this research are described in Section 4.4.

4.1. Speckle-like clutter cleaning

Since GAN models try to recreate the structure of the training data, it is important to remove these errors from the radar as suggested by Elsmann (2023), especially those with high dBZ values as the focus of this research is on high-intensity precipitation events. To highlight these values, density maps for exceeding 50 dBZ are generated for two full years to show where these reflectivity values occur. According to the Marshall-Palmer $Z - R$ relation, 50 dBZ corresponds to an intensity of 48.7 mm/h. The first year is 2008, under the clutter removal scheme described by Holleman and Beekhuis (2005), the second year is 2022, under the improved clutter removal scheme as described by Leijnse et al. (2016). This approach for showing the areas effected by clutter with high intensities was also performed by Van der Kooij (2021) and can be seen in Figure 4.2. In 2008 the areas with high density exceeding 50 dBZ closely represent the main shipping routes on the North Sea from Figure 4.1 as concluded by Van der Kooij (2021). In 2022 these ship tracks have lower densities, due to the improved clutter removal by the KNMI, but some areas with high reflectivity values align with the location of wind farms, also seen in Figure 4.1. These error sources can pose an issue for training precipitation nowcasting models on historic radar reflectivity data.

To remove the speckle-like clutter, a morphological clutter removal scheme is performed on each frame of the radar composite. First all pixels are converted to a wetmask, which is a binary map where 1 indicates rain rates exceeding 0.1 mm/h and 0 indicates rain rates lower than 0.1 mm/h. Then one or more erosion steps are performed where a boundary of the regions is removed from the wetmask. Then one or more dilation steps are performed where the boundaries are enlarged. This process of first applying erosion and then dilation is called opening. This method removes small areas with reflectivity which may be errors.

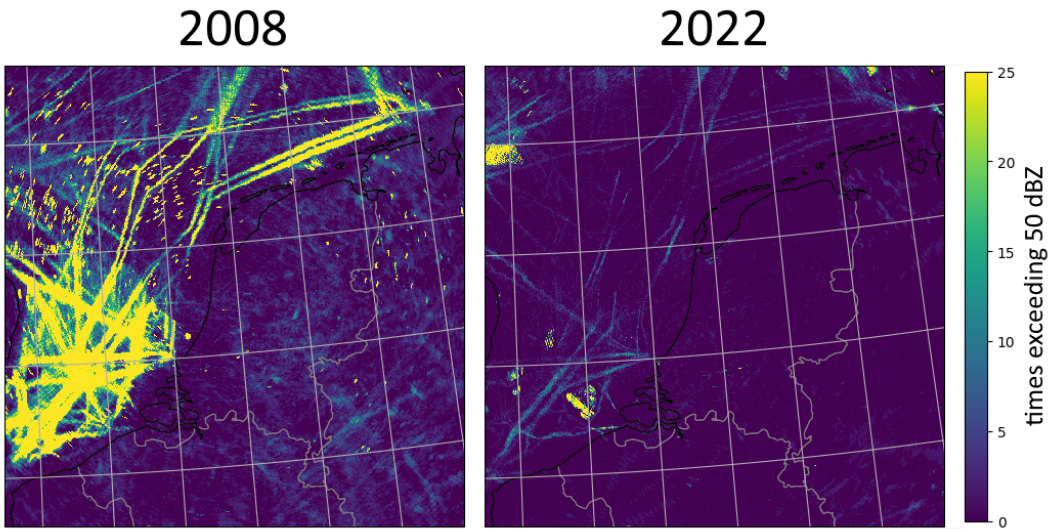

Figure 4.1.: Density maps for exceeding 50 dBZ for the years 2008 and 2022.

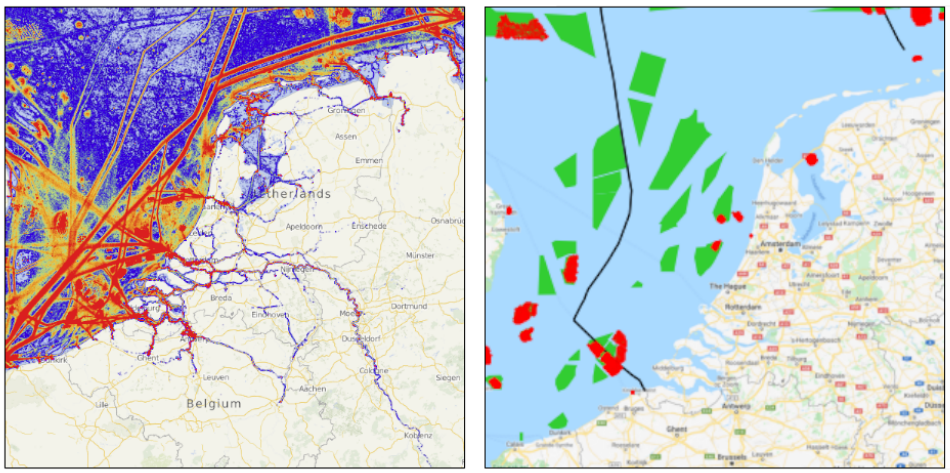

Figure 4.2.: Vessel density map, retrieved from www.vesselfinder.com (left); Wind farms in 2020, in red, and planned, in green, on the North Sea, retrieved from www.wins50.nl (right).

Errors within larger reflectivity areas are not removed with this method. An example of small precipitation fields with nearby clutter can be seen in Figure 4.3 and an example on large precipitation fields without clutter can be seen in Figure 4.4.

(a) Original. (b) Wetmask. (c) Erosion. (d) Dilation. (e) Cleaned.

Figure 4.3.: Morphological operations example on small precipitation fields with clutter.

(a) Original. (b) Wetmask. (c) Erosion. (d) Dilation. (e) Cleaned.

Figure 4.4.: Morphological operations example on large precipitation field without clutter.

Several settings for this method have been tried to find a balance between removing speckle-like errors from the data and not removing too much from the precipitation fields themselves. These settings vary with the number of erosion and dilation steps that are performed. These settings are indicated with two numbers, the first indicating the number of erosion steps that are performed first, followed by the number of dilation steps. The tested settings are: 2-2, 3-2, 3-3, 4-3 and 4-4.

These settings are evaluated by remaking the density figures for 2008 and 2022 after cleanup with these settings, as Figure 4.2, to indicate how much clutter with high reflectivity values is removed. Furthermore, histograms of the original and remaining intensity values are made for several events to give an indication of how much of the precipitation is removed. A balance between removing clutter and keeping high precipitation values is picked, 3-3, and applied to the full radar reflectivity dataset before further steps such as training the precipitation nowcasting models.

4.2. Data sampling method and application

First the event weighting method, in the form of theImportance Sampling Weight (ISW) is introduced in Section 4.2.1, followed by the application of these weights during training in Section 4.2.2 and ending with how incomplete events were handled in Section 4.2.3.

4.2.1. Importance Sampling Weight (ISW)

To reduce the number of examples containing little precipitation and to focus on high-intensity precipitation rates, a sampling strategy is devised. The model needs 22 consecutive frames (110 minutes) of 256×256, this will be defined as an event. A smaller research domain is selected which is used to determine the importance sampling weights, these can be seen in Figure 4.5. The model input is centered on De Bilt, based on the domain used by Elsmann (2023). The research domain starts 32 km from the edge of the model input as it cannot be expected that the model predicts what is coming from outside the model input domain. Due to the remaining clutter over the ocean after applying the clutter cleaning method from Section 4.1, the research domain is not taken over the ocean.

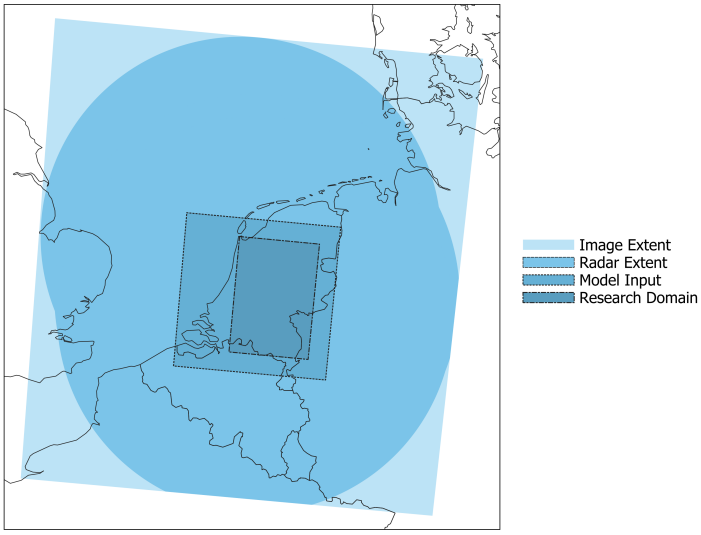

Figure 4.5.: The image extent of 700×765 in which the data is provided. The radar extent where there are radar measurements. The model input of 256×256 and the research domain of 134×192.

An event is taken and the precipitation rates of the cleaned data are $x_{n,c}$, where c indexes over $C = T \times h \times w$. The ISW is then determined using Equation 4.1 where P is the power parameter to tune the importance of precipitation rates higher than 1 mm/h in events. The weights were calculated for $P = 1.0$, $P = 1.5$ and $P = 2.0$.

$$ISW_e = \sum_C (x_{n,c})^P \tag{4.1}$$

Events are sorted based on their ISW and the highest 20% are used to train and validate the model to only select the events with precipitation within the research domain. In earlier work with the DGMR model, the importance sampling was used to randomly sample events to construct the training dataset where all events have a minimum probability of being selected, introducing events with no precipitation into the training data (Elsmann, 2023; Ravuri et al., 2021). With the new method, no random sampling is used when selecting the events within the dataset. The distribution of events with an ISW in the highest 20%

per month in the 2008-2022 dataset can be seen in Figure 4.6. This shows that there are more events selected in June, July and August. It can also be seen that with a higher power parameter more events are selected for May, June, July, August and September. It can also be seen that fewer events are selected in January, February, March, April, November and December with a higher power parameter.

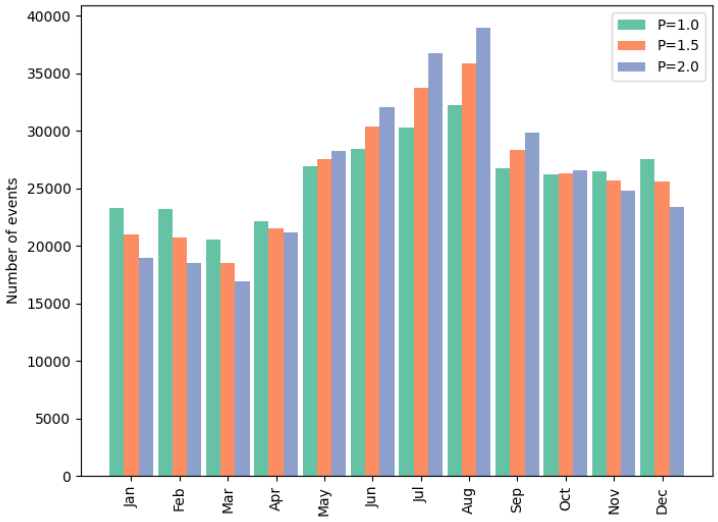

Figure 4.6.: The number of events per month in the 2008-2022 dataset after selecting the highest 20% for the different power parameter values.

4.2.2. Training strategies

Two training strategies are used, the first where the ISW is used as a threshold to sample the events on during the training with replacement, further described as the unweighted method. All events have an equal probability of being selected during the training with this threshold method and they can occur multiple times. The threshold will be the top 20% of weights with $P = 2.0$. $P = 2.0$ is selected as this emphasises events with higher precipitation rates, especially high-intensity precipitation rates. The second training method uses the ISW to determine the probability of selecting an event during training with Equation 4.2, further called the weighted method. Here the sum is only taken over the highest m weights, within the top 20% of weights to ensure that the sum of all probabilities add up to 1. In this weighted method the events are again drawn with replacement and the training is focused more on events with higher weights where higher precipitation intensities occur. This focus can be seen in Figure 4.7 showing the probability of selecting an event, where the events are sorted on the ISW. Due to the higher probability of selecting events with a high weight, $P = 2.0$ is used as the ISW during training.

$$q_e = \frac{ISW_e}{\sum_m ISW} \tag{4.2}$$

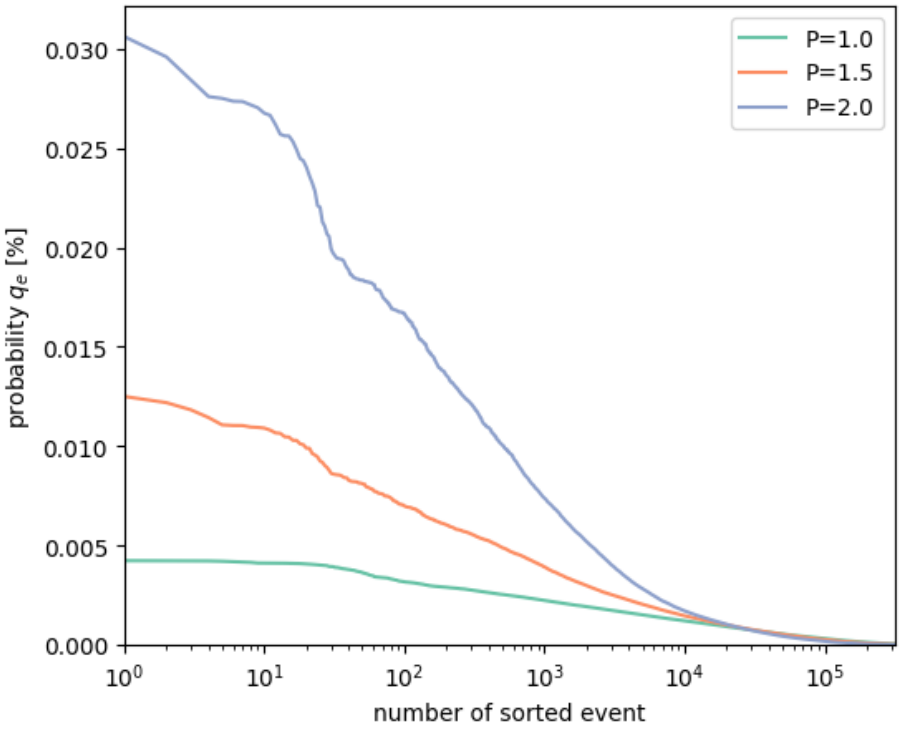

Figure 4.7.: The probability of selecting event e, q_e, for the top 20% events in the 2008-2022 dataset for the different power parameter values. The events are sorted on the ISW.

4.2.3. Incomplete events

The single image provided as the reflectivity composite on the data platform of the KNMI is a combination of the data of two radar systems. To ensure that this is homogeneous in the training data, a check is performed on all time steps to ensure that the radar reflectivity is available and is covered by two radar systems, one at Den Helder and another at either De Bilt or Herwijnen. Out of all possible events in the dataset, 1,577,931 events, this restriction removes 118,243 events, or around 7.5%. The code base of DGMR is based on the work from Elsmann (2023), which also included a model version where the Echo Top Height (ETH) product of the KNMI was used. Due to this the ETH dataset was also checked on the availability and coverage, removing a further 11,496 events, or 0.7%. More information about this product and an analysis of precipitation rates and ETH can be read in Appendix E. Events with one or more frames without full coverage from both radar products are assigned an ISW of zero and will therefor not be used during training.

4.3. DGMR model training

First, the processing for the training data is explained, including the training, validation and test split of the dataset where the distribution of the events is shown in Section 4.3.1. After this, the changes to the DGMR model as well as the settings for the experiments are shown in Section 4.3.2.

4.3.1. Data preparation

After the weights have been calculated for all valid events and the top 20% have been selected, 315,591 events remain in the dataset where Figure 4.8 shows how many there are per year. The dataset has to be split into a training, validation and test set. To prevent many consecutive events, which are closely correlated to each other, being split among these three sets, the training set will consist of all events from 2008 up to and including 2020, the validation set has all events from 2021 and the test set 2022. This results in 279,766 training events, 18,822 validation events and 17,003 test events, the distribution per month being shown in Figure 4.9.

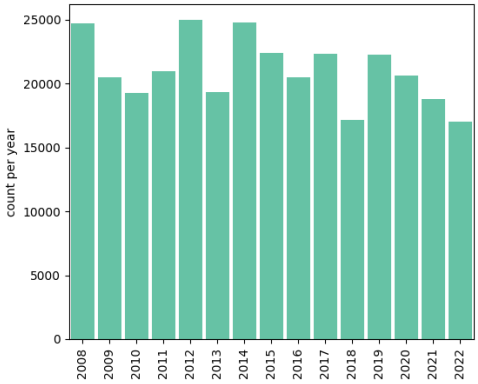

Figure 4.8.: The number of top 20% events per year in the dataset.

In the original work with this version of the DGMR model from Elsmann (2023), the Precipitation data as well as ETH data, see Appendix E, were saved in TFRecord files where one file contained the data of a single day. However, this has two limitations, the first being that valid events that have frames in two days have not been included, leaving out all events with the first input frame, t_{-20}, at 22:15 UTC up to 00:00 UTC, being about 7.3% of all valid events. Second of all, this storage method leaves poor control over the probability of selecting each event individually during training as each TFRecord file may contain up to 267 events. Therefore, the precipitation data is stored in TFRecords, where each file is a valid event with a weight in the top 20% since only these will be used during training. While producing the TFRecord files, all pixels with an precipitation rate over 200 mm/h where set to 200 mm/h. Values higher than 200 mm/h may be erroneous in nature or indicate hail, this reduces some extreme values, some even over 1.000 mm/h as the model will be penalized significantly when it is unable to predict these extremes.

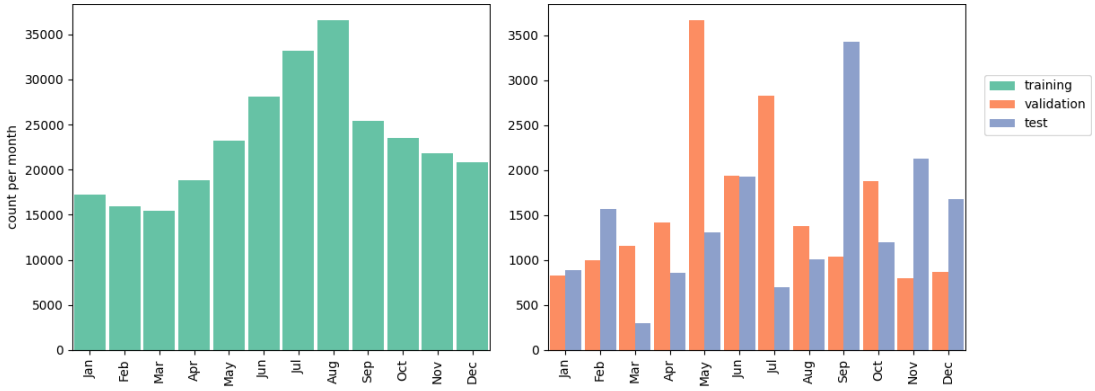

Figure 4.9.: The number of top 20% events per month for the training set on the left and the validation and test set on the right.

4.3.2. Training setup

Two experiments are done on DelftBlue (Delft High Performance Computing Centre , DHPC) using one NVIDIA Tesla V100S 32GB for each experiment due to the limited number of GPU's. Training may take up to 120 hours, after which it is automatically stopped to ensure the limited number of available GPU's can be used by other users. This severely limits the resources and time required to train the full model to a stable state. To ensure the training reaches stability with one GPU and the time limitation, the blocks have been reduced in size in the generator as well as the discriminators. This is done by reducing the size of the convolution layers in the generator and the down-sampling blocks in the discriminators, see Figures 4.10 and 4.11 for an overview. The original layout of the model can be seen in Appendix D.

The training was done using TensorFlow (Adabi et al., 2016) 2.8.2 with a batch size of 16 and 100.000 training steps. The model takes the previous four radar observations, the previous 20 minutes, as context and forecasts for the next 18 frames, the next 90 minutes. The generator and discriminator make use of the Adam optimizer, which was set to a learning rate of 5×10^{-5} and 2×10^{-4} respectively as was done by Elsmann (2023). The discriminators are updated twice for every training step compared to once for the generator. Every 500 training steps, the model performs 100 validation steps to track the training progress. Both experiments, one with the weighted selection scheme and another with the unweighted selection scheme, as described in Section 4.2.2, are performed with the other settings left the same.

4.4. Model verification

The forecasts of the two training strategies described in Section 4.2.2 are evaluated against each other using metrics calculated on 1000 randomly selected events from the test set, which are described in Sections 4.4.1 and 4.4.2. An adaptation of a metric used for determining a models capability with forecasting the peak discharge in a catchment is introduced in the form of the Peak Anticipation Time in Section 4.4.3. All metrics are calculated on the research domain as described in Section 4.2.1. The predictions are also compared visually on a subselection of test events against the predictions from S-PROG. A description of the

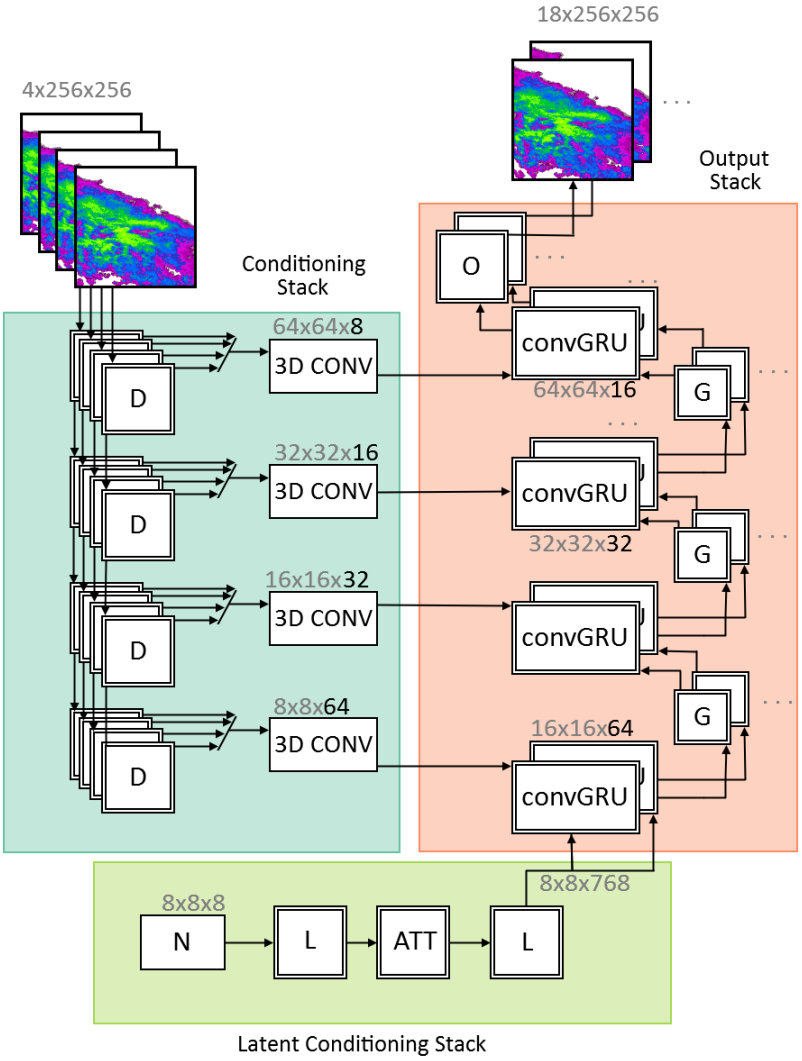

Figure 4.10.: Schematic overview of the DGMR Generator with the Conditioning Stack where the 4 input frames of size 256 × 256 are processed, the Latent Conditioning Stack where noise from a Gaussian distribution is processed and fed towards the Output Stack where 18 output frames of size 256 × 256 are generated. Image taken from Elsmann (2023) with further details of the blocks in Figure 4.11 and where changes to the size of blocks are indicated in bolt.

selection of these test events is provided in Section 4.4.4. S-PROG was setup to take 4 input frames and give 18 output frames to compare it with similar information as input and for the same lead time as the generative model.

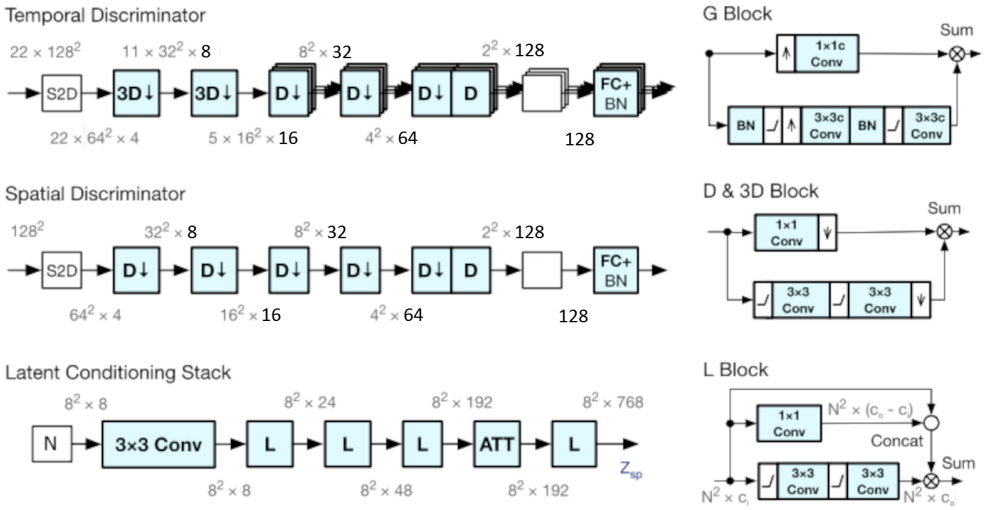

Figure 4.11.: Discriminators, Latent Stack and the architecture of the Generator block, Downsampling block and Latent block used in DGMR, where changes to the size of blocks in the Discriminators are indicated in bolt. Image taken from Ravuri et al. (2021).

4.4.1. Continuous metrics

Two continuous scores are calculated as a function of lead time. These are the Mean Absolute Error (MAE) and Mean Squared Error (MSE).

$$MAE = \frac{1}{N \times 134 \times 192} \sum_{n=1}^{N} \sum_{i=1}^{134} \sum_{j=1}^{192} |F_{i,j} - O_{i,j}| \tag{4.3}$$

$$MSE = \frac{1}{N \times 134 \times 192} \sum_{n=1}^{N} \sum_{i=1}^{134} \sum_{j=1}^{192} (F_{i,j} - O_{i,j})^2 \tag{4.4}$$

Where N is the number of samples, F the forecast and O the observed precipitation and the area of the research domain is 134 by 192. The MSE puts more emphasis on larger errors. Both metrics range from 0, a perfect score, to infinity. These metrics have been calculated on the precipitation intensities in mm/h.

4.4.2. Categorical metrics

Categorical metrics are used to give a qualitative assessment of the forecast for different precipitation thresholds. The metrics used are Critical Success Index (CSI), Probability of Detection (POD), False Alarm Rate (FAR), F1 score and Fraction Skill Score (FSS). The forecasts and observations are converted into binary maps depending on whether the pixel values are above or below the threshold intensity values. The categorical metrics are computed based on the four elements of the confusion matrix in Table 4.1. The considered intensity thresholds are 1 mm/h, 10 mm/h and 20 mm/h.

Table 4.1.: The confusion matrix outcomes.

		Observation	
		Positive	Negative
Forecast	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Critical Success Index (CSI) indicates how well the true positives correspond to the total number of cases minus the true negative. It ranges from 0 to 1, where higher is better.

$$CSI = \frac{TP}{TP + FP + FN} \tag{4.5}$$

Probability of Detection (POD) indicates how well the true positives correspond to the observed positives. It ranges from 0 to 1, where higher is better and overpredictions tend to give a better score.

$$POD = \frac{TP}{TP + FN} \tag{4.6}$$

False Alarm Rate (FAR) indicates how many false positives are given as a fraction of the total forecasted positives. It ranges from 0 to 1, where lower is better. Overpredictions will give a worse score by increasing the FAR.

$$FAR = \frac{FP}{TP + FP} \tag{4.7}$$

F1 indicates the balance between the precision, how many forecasted positives are true positive, and the recall, how many observed positives were correctly predicted (TP). It ranges from 0 to 1, where higher is better.

$$F1 = \frac{TP}{TP + 0.5 \times (FP + FN)} \tag{4.8}$$

Fraction Skill Score (FSS) is a spatial verification score, which indicates the skill of a forecast to predict above a given precipitation thresholds and for different spatial scales of size n. It is calculated over lead time and ranges from 0 to 1, where higher is better (Roberts and Lean, 2008). The average FSS over all samples per scale and lead time is calculated.

$$FSS = 1 - \frac{MSE_{(n)}}{MSE_{(n)ref}} \tag{4.9}$$

with

$$MSE_{(n)} = \frac{1}{N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} (O_{(n)i,j} - F_{(n)i,j})^2 \tag{4.10}$$

$$MSE_{(n)ref} = \frac{1}{N_x N_y} \left(\sum_{i=1}^{N_x} \sum_{j=1}^{N_y} O^2_{(n)i,j} + \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} F^2_{(n)i,j} \right) \tag{4.11}$$

where N_x and N_y are the number of columns and rows in the radar data, i and j indicate the row and column index of the fraction matrices respectively, and $O_{(n)i,j}$ and $F_{(n)i,j}$ are the fraction of pixels exceeding the threshold precipitation intensity in the fraction matrix i, j of the observation and forecast respectively.

The FSS is determined for windows of size 4 km, 8 km, 16 km and 32 km.

4.4.3. Peak Anticipation Time score

The Peak Anticipation Time (PAT), originally proposed by Imhoff et al. (2022) in the context of hydrology, is the first issue time for which the maximum discharge was forecasted within a given magnitude range. In this research, the PAT is applied to an image and will indicate how often the peak precipitation within a window of a given size is forecasted within the window and within a given magnitude range, for any given lead time. This method can be applied to ensemble predictions or on deterministic predictions. It ranges from 0 to 1, where higher is better.

For window i, j of size $n \times n$ it is determined if the maximum observed precipitation rate, $O_{(n)i,j}$, is above the threshold value. Then the remaining windows are assigned a 1 if the maximum forecasted precipitation rate, $F_{(n)i,j}$, is within a factor of tolerance, f, and 0 otherwise.

$$W_{(n)i,j} = \begin{cases} 1, & \max\left(F_{(n)i,j}\right) \times f \geq \max\left(O_{(n)i,j}\right) \geq \frac{\max\left(F_{(n)i,j}\right)}{f} \\ 0, & \text{else} \end{cases} \tag{4.12}$$

The PAT is then calculated by taking the average of the values assigned to the windows that have a maximum observed precipitation rate above the threshold value, $W_{(n)i,j}$.

$$PAT = \overline{W}_{(n)i,j} \tag{4.13}$$

The time aspect of the PAT is introduced by determining it for a single time step but with different lead time, a simple representation of this is shown in Figure 4.12.

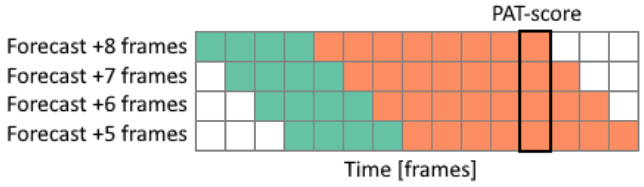

Figure 4.12.: A representation of the PAT-score being determined over lead time on a model with 4 input frames and 8 output frames. The columns indicate the time in frames and each row indicates how the PAT is determined for a given lead time in a number of frames.

The PAT-score is determined on a subselection of test events with a factor of tolerance of 1.5, with intensity thresholds of 1 mm/h, 10 mm/h and 20 mm/h and for windows of size 4 km, 8 km, 16 km and 32 km. In part due to the random noise introduced in every forecast generated by the generative model, forecasts can differ significantly when looking at the same timestep. To reduce the random noise of the model influencing the PAT-score significantly between two consecutive predictions, an ensemble of 10 forecasts is used to determine the PAT-score also averaging the PAT-score over the ensemble.

4.4.4. Test events verification

For the qualitative assessment of the models and for determining the PAT-score of the trained models, 20 events were selected from the test set by looking for the highest ISW using $P = 2.0$ as described in Section 4.2.1 to have the emphasis on events with high precipitation intensities. Since events starting 5 minutes apart have a lot of overlap, their ISW will be very similar. To make sure that each event is unique, only the event with the highest ISW is selected within a two hour time frame. These 20 events have some variance in the behaviour of the precipitation field and similar events were removed as to focus on the behaviour of the model under completely different atmospheric conditions, reducing it to only 6 unique test events. These can be seen in Figure 4.13. This selection was done before the model training was completed.

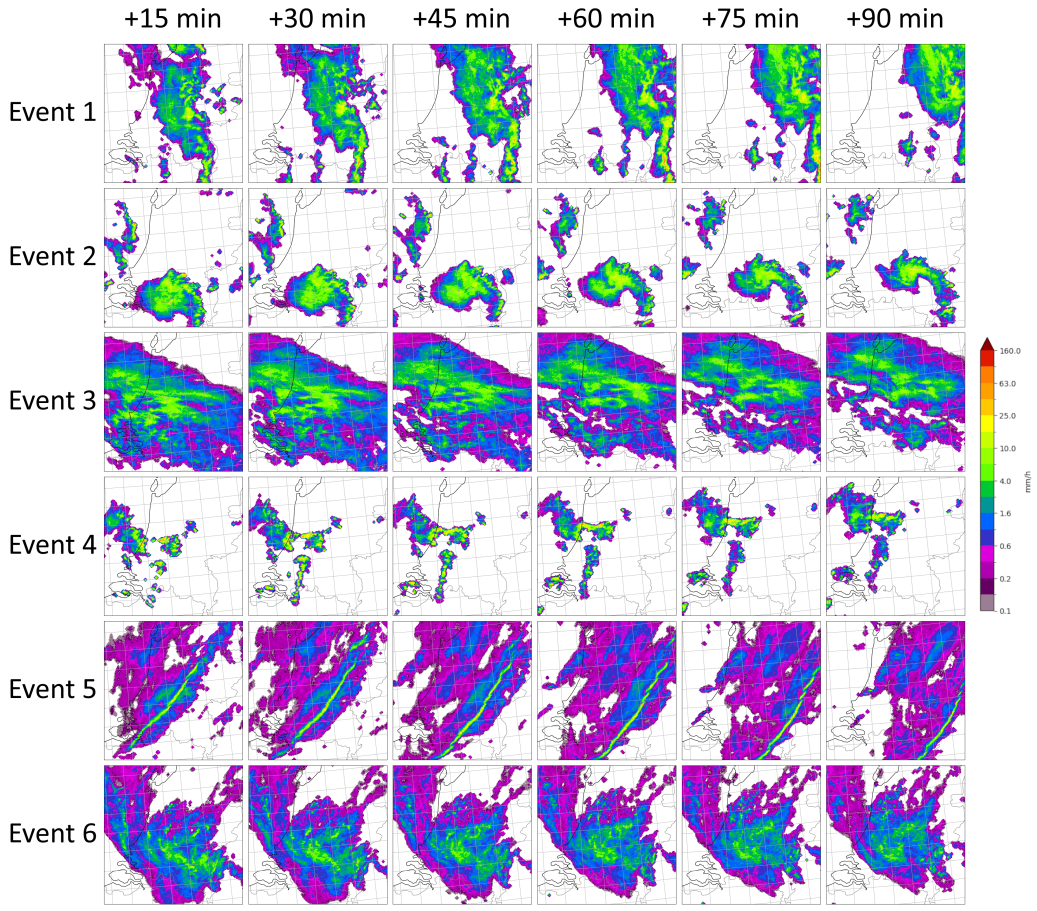

Figure 4.13.: The 6 unique test events used for qualitative assessment.

For all events the given time and date indicate the last of four input frames, at $t = 0$ min and a short description of each event is focused on the behaviour within the research domain.

Event 1: 2022-05-19 12:05 A large precipitation field moving towards the Northeast with little change to its shape. The maximum precipitation intensities are between 20 and 60 mm/h but drop to just above 10 mm/h at end of the prediction and the median is approximately 2 mm/h for the entire duration.

Event 2: 2022-09-08 20:25 A precipitation field covering up to a quarter of the research domain moving towards the Northeast spreading in Northwest-Southeast direction. The peak precipitation intensities are between 20 and 100 mm/h for some cells within the larger precipitation field which may be of convective nature and the median is approximately 3 mm/h at the start and drops to 2 mm/h at the end of the event.

Event 3: 2022-12-23 12:20 A large precipitation field covering nearly the entire research domain and moving towards the East and slightly to the North with barely any change in shape of the precipitation field. The peak intensities are between 10 and 20 mm/h with a median of 1 mm/h.

Event 4: 2022-08-17 01:55 Some smaller precipitation fields forming at the start of the event and moving North. These are very likely formed due to convection and the peak intensities are between 40 and 100 mm/h within the small convective cells. The median precipitation intensity is around 1.5 mm/h and drops to 1 mm/h from t_{+60} minutes.

Event 5: 2022-02-20 20:55 A large precipitation field with high-intensity precipitation in a line structure moving towards the East. This line structure was formed by a cold front moving in from the West, giving peak precipitation intensities of 25 up to 52 mm/h while the median precipitation rate was close to 0.5 mm/h for the entire duration.

Event 6: 2022-06-05 15:00 A large precipitation field covering nearly 80% of the research domain moving slowly towards the North with an anti-clockwise rotation. The peak precipitation intensities are between 15 and 20 mm/h and drop to just below 10 mm/h at the end of the event and the median intensity is around 1 mm/h.

5

Results

First, the effect of the clutter cleanup method under different settings is analysed in Section 5.1. This analysis uses the density maps of 2008 and 2022 and examines the effect on two example radar images throughout the training dataset. Next, the performances of the trained models are analysed in Section 5.2. The average metrics from 1000 randomly sampled events, uniformly selected from the top 20% of weights in the test set, are presented. The models are then visually compared using the forecasts on six test events, using S-PROG as a benchmark model. Finally, the performance of the two trained DGMR models regarding their predictive capability of peak intensities over lead time is analysed using the Peak Anticipation Time (PAT)-score on these six events.

5.1. Speckle-like clutter cleanup

The effect of different settings for the speckle-like clutter cleaning is evaluated by comparing the original echo density plots for values above 50 dBZ in 2008 and 2022 to the density plots after speckle-like clutter removal under various clutter removal settings in Section 5.1.1. To demonstrate the impact of this clutter removal strategy on individual frames and its effect on data removal in precipitation fields, two frames are shown in Section 5.1.2. The first with speckle-like clutter and the second with several precipitation fields of different sizes.

5.1.1. Density maps

The density plots for 2008 and 2022 after cleanup are shown Figures 5.1 and 5.2, respectively. In the 2008 plot, the density of the ship tracks is noticeably reduced for all settings, with the 4-3 strategy showing the most reduction. However, remnants of ship tracks persist in certain areas, particularly over the sea in the North-West, forming a ring-like pattern centered around the Den Helder radar. Additionally, high reflectivity pixels persist off the coast of Rotterdam under all settings, likely due to ground clutter at fixed locations. The presence of residual clutter after cleanup suggests its possibly embedded within precipitation fields. Furthermore, land areas with a high density of reflectivity exceeding 50 dBZ are entirely cleared for all settings except 2-2. Remaining high reflectivity occurrences over land may be due to precipitation events or other error sources without a fixed location.

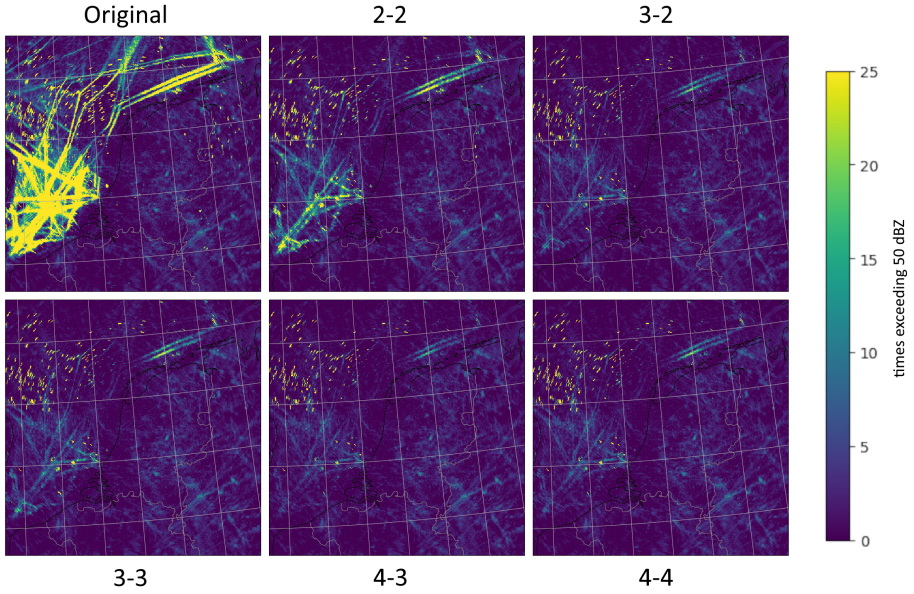

Figure 5.1.: Density maps for exceeding 50 dBZ for the year 2008 in the original dataset and after applying the different speckle-like clutter cleanup settings.

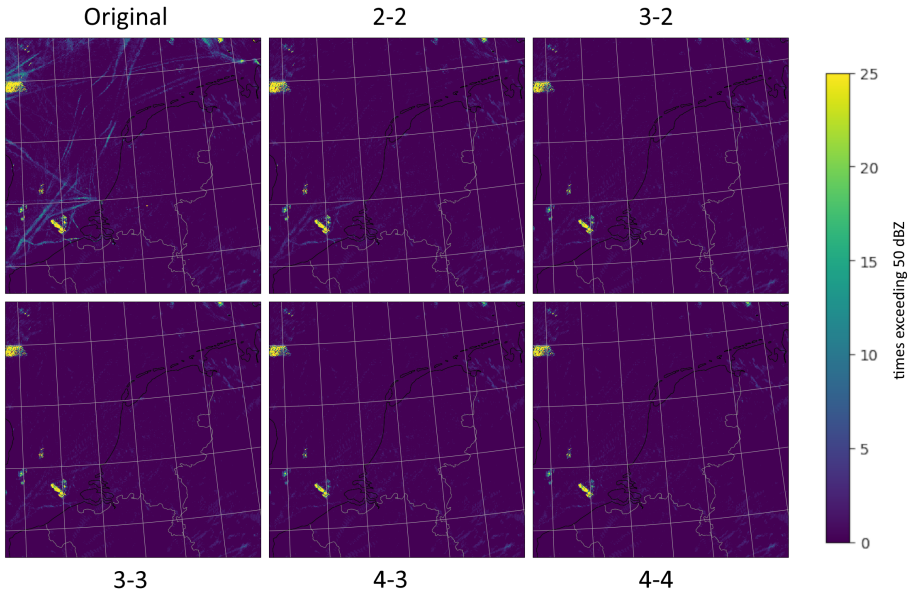

Figure 5.2.: Density maps for exceeding 50 dBZ for the year 2022 in the original dataset and after applying the different speckle-like clutter cleanup settings.

For 2022, some ship tracks remained visible in the original dataset. These are effectively removed when three or more erosion steps are performed, with exceptions in the corners of the image. The high reflectivities caused by wind warms persist after cleanup, suggesting

that these occur within larger fields. Similarly, some high reflectivity occurres over land and are not removed, likely due to them occuring within larger fields as well. These could represent true precipitation of high intensity or errors resulting from the processing method applied by KNMI. The density maps illustrate that increasing the number of erosion steps removes more high reflectivity values from the data. They also show that fewer dilation steps result in the removal of more high reflectivity values.

5.1.2. Illustration of clutter removal on example radar images

The clutter removal technique is evaluated on two examples. The first example, Figure 5.3, contains speckle-like clutter without precipitation. In this situation, the clutter filter successfully removes over 95% of all the clutter, regardless of the chosen settings, see Figure 5.4 for more details. In the second example, Figure 5.5, both large and small precipitation fields are present. Here over 50% of the lowest intensity is removed with the 3-2 and 4-3 settings, see Figure 5.6 for more details. Since speckle-like clutter is absent in this example, a lower percentage of higher intensities is removed, as these high precipitation rates occur within the precipitation fields away from the edge. More examples can be seen in Appendix A.

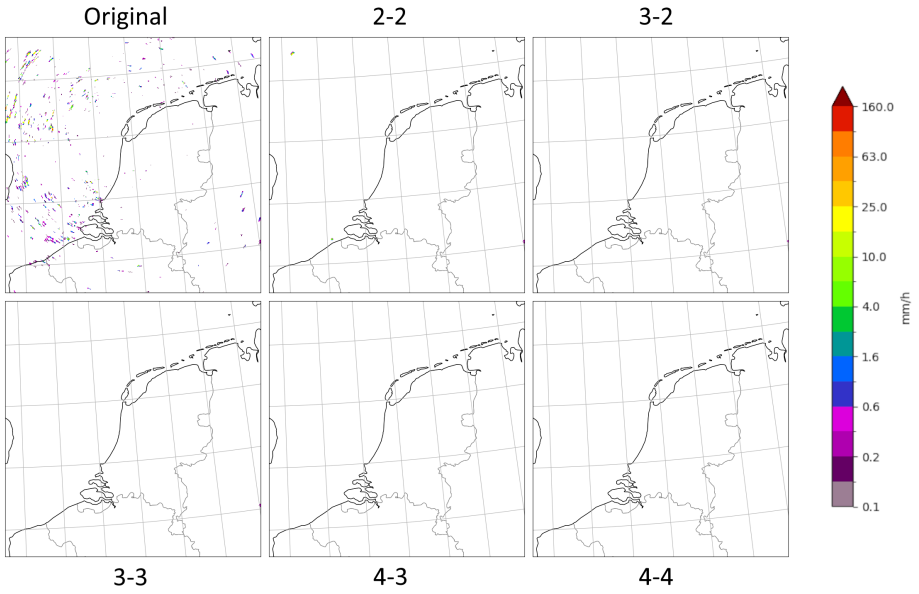

Figure 5.3.: Example of a radar image before and after cleanup under different settings taken on July 2nd, 2009 at 10:00 UTC containing only speckle-like clutter.

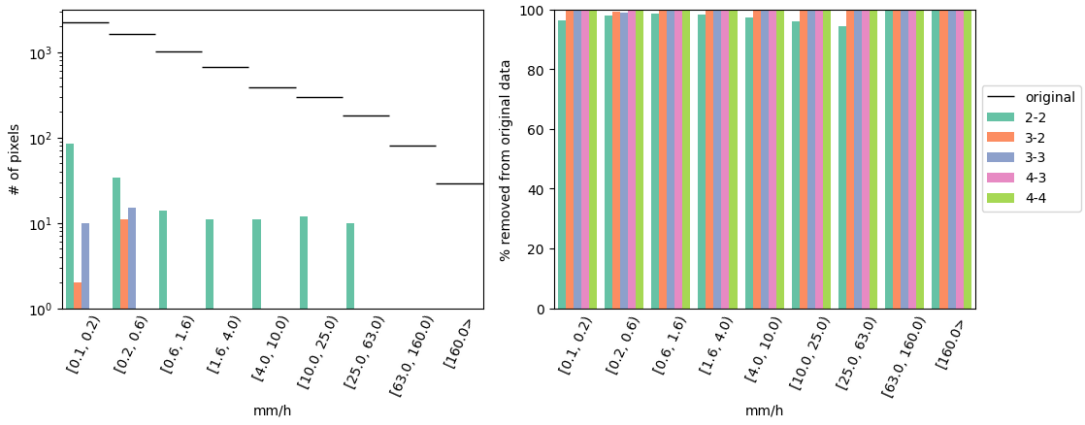

Figure 5.4.: Extra detail for Figure 5.3. The number of pixels per intensity bin after cleaning with the different settings compared to the original number of pixels (left). The number of pixels per intensity bin after cleaning with the different settings compared to the original number of pixels (right).

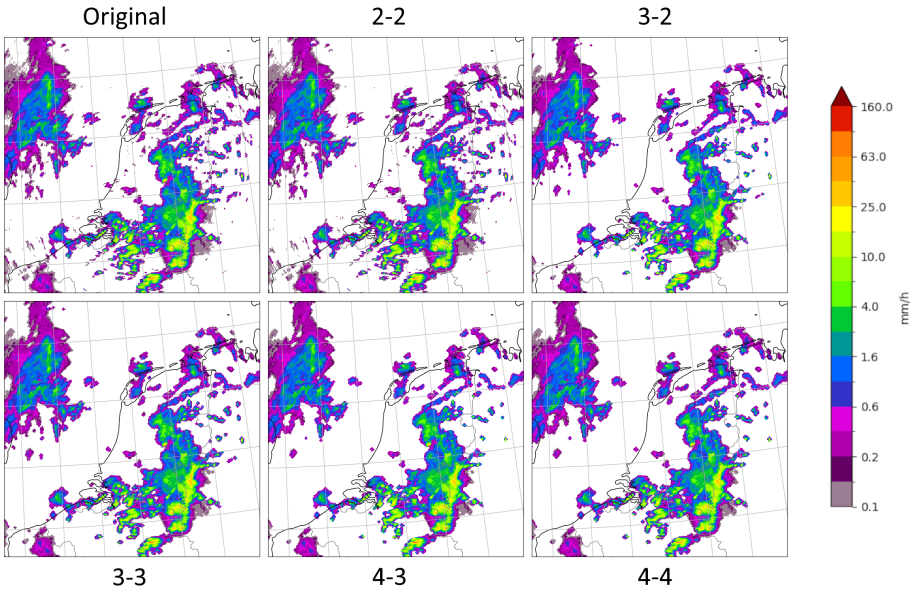

Figure 5.5.: Example of a radar image before and after cleanup under different settings taken on August 18th, 2011 at 17:00 UTC with several large and small precipitation fields with little to no clutter.

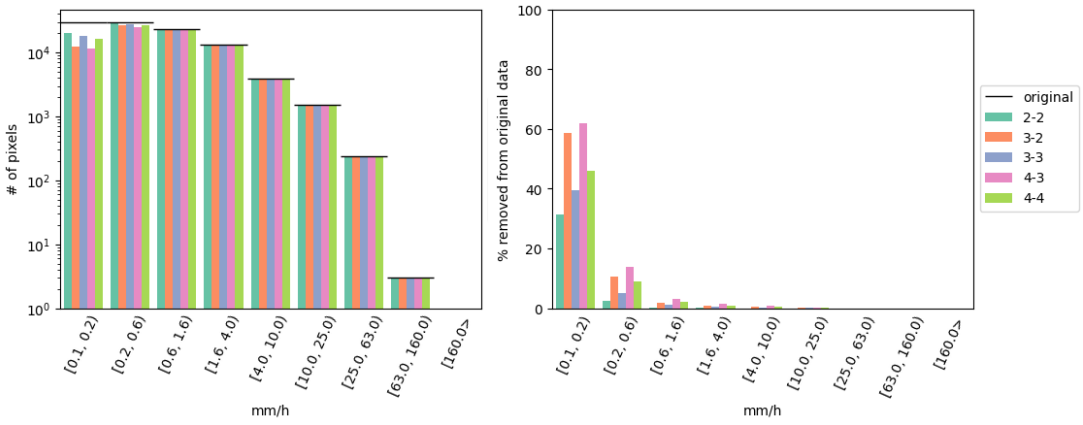

Figure 5.6.: Extra detail for Figure 5.5. The number of pixels per intensity bin after cleaning with the different settings compared to the original number of pixels (left). The number of pixels per intensity bin after cleaning with the different settings compared to the original number of pixels (right).

5.2. Nowcasting model results

The results from the model runs will be presented in two parts. The first part will contain the commonly used metrics. The second part will contain the results from the forecasts made on the test events, including the predictions made by S-PROG, along with the PAT-score of the DGMR models for each event.

5.2.1. Metric results

Table 5.1 indicates poor performance by both models across all pixel-wise metrics, particularly for intensities $\geq$ 10 mm/h and $\geq$ 20 mm/h. The weighted model demonstrates slightly improved average performance on CSI, F1 and a larger improvement on POD for all intensities. The unweighted model shows slightly improved performance on average for FAR. This is attributed to the weighted model predicting higher intensities, resulting in improved scores for CSI, POD and F1, but a worse result for FAR due to overestimation of the area with these precipitation rates compared to the observations. Consequently, the POD significantly increased with the weighted model compared the the unweighted model due to overprediction.

Table 5.1.: Average categorical metrics on 1000 randomly sampled events from the test set for different threshold values on the DGMR model under the two different training strategies. Where (>) indicates higher values are better scores and (<) indicates that lower values are better.

	CSI (>)			POD (>)		
	$\geq$ 1 mm/h	$\geq$ 10 mm/h	$\geq$ 20 mm/h	$\geq$ 1 mm/h	$\geq$ 10 mm/h	$\geq$ 20 mm/h
Unweighted	0.106	0.007	0.002	0.416	0.034	0.014
Weighted	0.113	0.008	0.003	0.528	0.065	0.024
	FAR (<)			F1 (>)		
	$\geq$ 1 mm/h	$\geq$ 10 mm/h	$\geq$ 20 mm/h	$\geq$ 1 mm/h	$\geq$ 10 mm/h	$\geq$ 20 mm/h
Unweighted	0.858	0.979	0.987	0.177	0.013	0.004
Weighted	0.861	0.988	0.992	0.188	0.015	0.006

From Figure 5.7, it can be seen that the MAE and MSE for the unweighted model is lower. Both MAE values increase rapidly up to t_{+35} and t_{+45} for the unweighted and weighted models, respectively, after which they stabilize. The MSE for both models increases rapidly before decreasing for the unweighted model. The higher MAE and MSE values for the weighted model are attributed to its tendency to predict higher precipitation rates over larger areas, resulting in poorer results pixel-wise statistics statistics on average.

Figure 5.8 illustrates that the FSS decreases with a decreasing window scale for both models and increasing intensity threshold. moreover, the weighted model tends to predict higher precipitation rates on average, as evidenced by the higher FSS values for $\geq$ 10 mm/h and $\geq$ 20 mm/h across all lead times, except t_{+65} for $\geq$ 20 mm/h. However, the unweighted model performs slightly better for $\geq$ 1 mm/h at lead times longer than t_{+40} on all scales, owing to the weighted model's tendency to overestimating lower intensities.

For an overview of model performance on these metrics for the top 1%, top 2%, top 5% and top 10% events selected in the test set, see Appendix B.

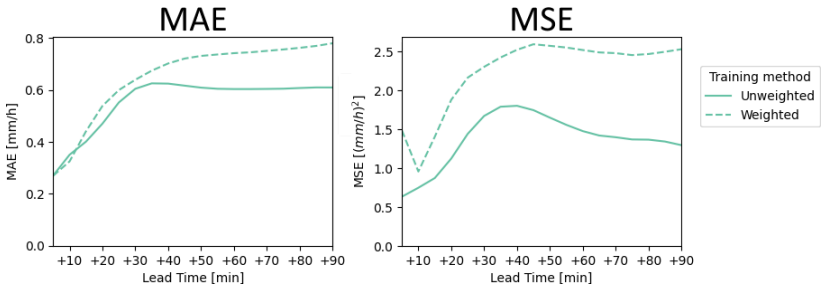

Figure 5.7.: Average Mean Absolute Error (MAE) and Mean Squared Error (MSE) on 1000 randomly sampled events from the test set on the DGMR model under the two different training strategies.

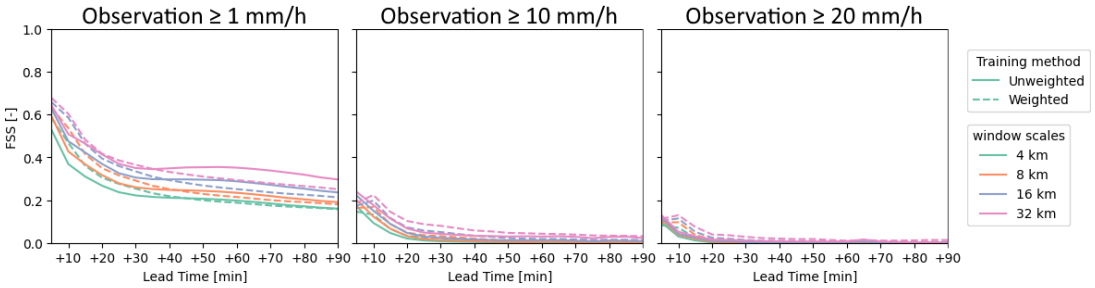

Figure 5.8.: Average Fraction Skill Score (FSS) on 1000 randomly sampled events from the test set for different threshold values on the DGMR model under the two different training strategies, higher is better.

5.2.2. Test event forecasts

In the following Section, the forecasts for individual events, as described in Section 4.4.4, are shown along with their scores. Additionally, all events are plotted on a linear scale, which can be found in Appendix C.

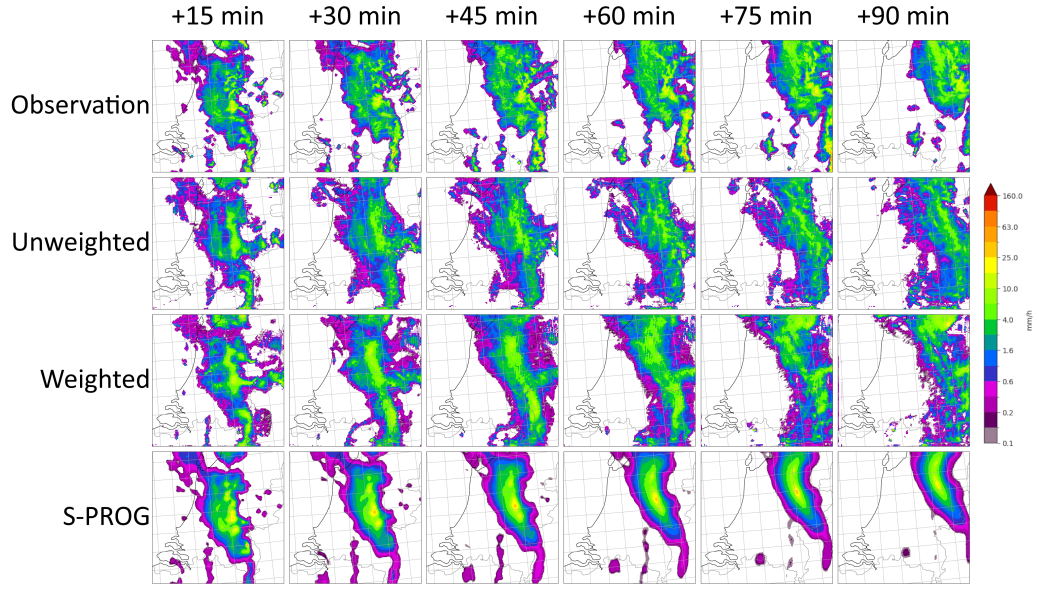

Figure 5.9.: Observation and unweighted and weighted DGMR and S-PROG predictions for Event 1, at t_0 2022-05-19 12:05 UTC.

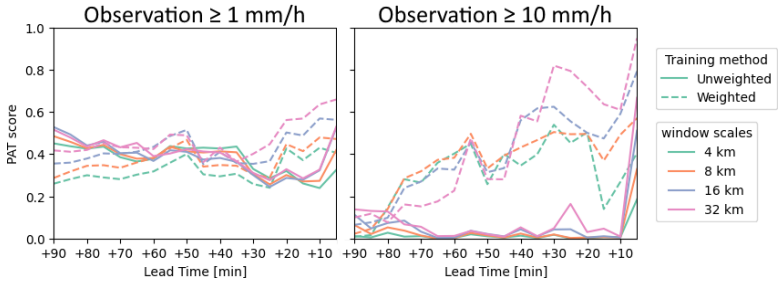

Figure 5.10.: The PAT-score for Event 1 of an ensemble of 10 of each model, at t_{+90} where there is no observed precipitation intensity of $\geq$ 10 mm/h.

Forecast Event 1

From Figure 5.9, it's evident that S-PROG forecasts gradually adopt a more generic and rounded shape with lead time, smoothing and fading the precipitation fields. While the unweighted model closely resembles the observation, it incorrectly predicts Northeast motion as directly Eastward. This motion discrepancy is also present in the weighted model, suggesting that the model has learned the average wind direction of the training data rather than accurately predicting the motion field of individual events.

Figure 5.10's PAT-score of the weighted and unweighted DGMR models show both capture peak intensities $\geq$ 1 mm/h from the start, yielding higher scores with larger scales. However, the unweighted model outperforms the weighted model until t_{+30}. For intensities $\geq$ 10 mm/h, the weighted model significantly outperforms the unweighted model, demonstrating comparable performance for high intensities ($\geq$ 10 mm/h) compared to intensities $\geq$ 1 mm/h.

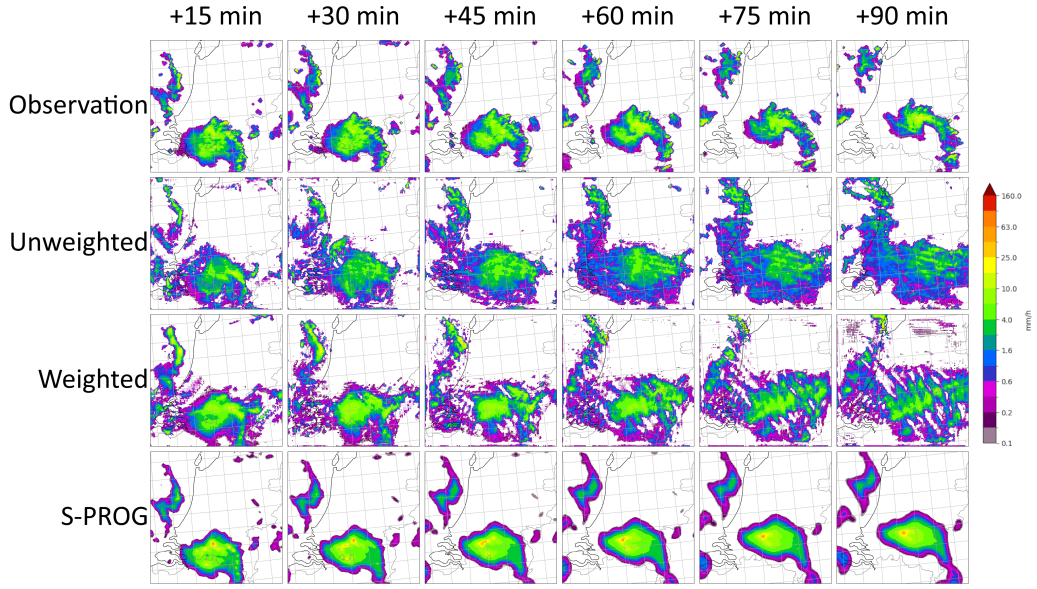

Figure 5.11.: Observation and unweighted and weighted DGMR and S-PROG predictions for Event 2, at t_0 2022-09-08 20:25 UTC.

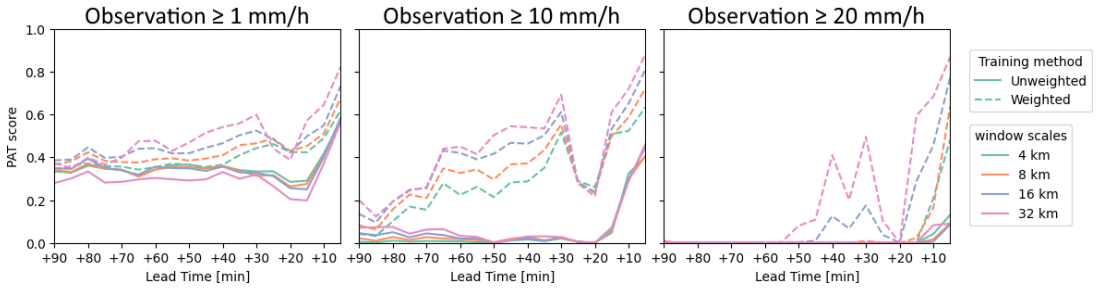

Figure 5.12.: The PAT-score for Event 2 of an ensemble of 10 of each model, at t_{+90}.

Forecast Event 2

In Figure 5.11, S-PROG forecasts excel at longer lead times, with both unweighted and weighted DGMR models expanding the precipitation field but failing to capture high-intensity areas effectively. This expansion aligns with findings in Elsmann (2023) with the DGMR model. A parallel line structure emerges, particularly in the weighted model, when the observed motion field deviates significantly from the average wind direction.

Figure 5.12's PAT-score reflects these results, with both models performing well for intensities ≥ 1 mm/h, but the weighted model significantly outperforms the unweighted model for intensities ≥ 10 mm/h. The weighted model also predicts precipitation rates ≥ 20 mm/h within a factor of 1.5 of the observed rate from t_{+50} at larger window scales. However, there's a performance drop at t_{+25} and t_{+20}, likely due to fewer cells with high precipitation rates in the input frames at these lead times, leading to inaccurate predictions.

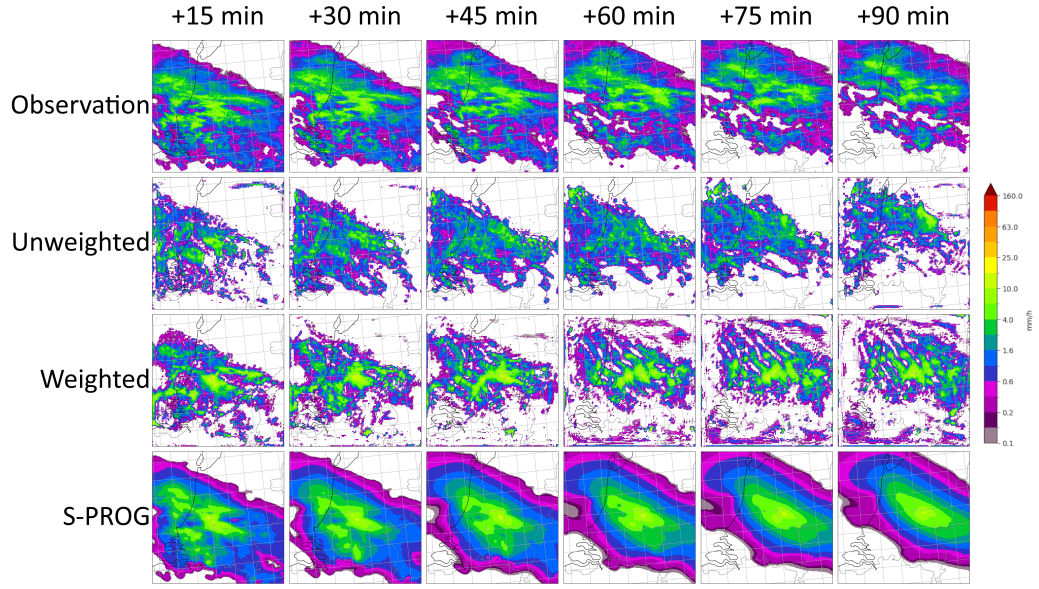

Figure 5.13.: Observation and unweighted and weighted DGMR and S-PROG predictions for Event 3, at t_0 2022-12-23 12:20 UTC.

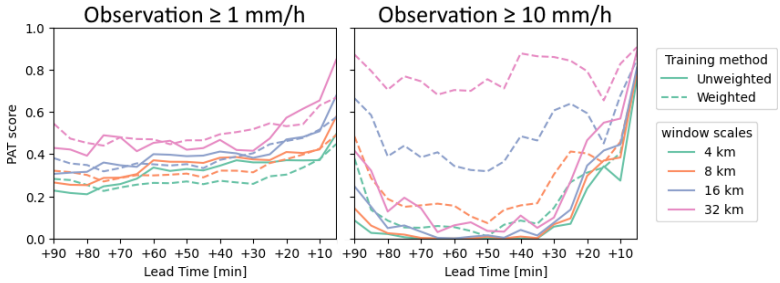

Figure 5.14.: The PAT-score for Event 3 of an ensemble of 10 of each model, at t_{+90} where there is no observed precipitation intensity of $\geq$ 20 mm/h.

Forecast Event 3

In Figure 5.13, S-PROG exhibits the best forecasted motion, with the precipitation field moving towards the Northeast. However, in both the unweighted and weighted models, a parallel line structure is observed, resulting in poor results for this event. Figure 5.14's PAT-score remains around 0.4 for both models when observations are $\geq$ 1 mm/h, with a slight increase with shorter lead times. Initially, both models correctly forecast some intensities $\geq$ 10 mm/h, particularly with the 32 km window scale in the weighted model, averaging around 0.8 for every lead time. However, the PAT-score of the unweighted model and smaller window scales of the weighted model drop immediately and begin to rise again around t_{+35}. This drop is attributed not to lower precipitation rates in the input frames, but to the formation of the parallel lines extending higher precipitation rates further North and East due to the forecast motion, which aligns with the observed motion towards the Northeast. The weighted model's quicker formation of lines, as well as the higher forecasted intensities, enable it to predict peak intensities where $\geq$ 10 mm/h occur in the observation more frequently.

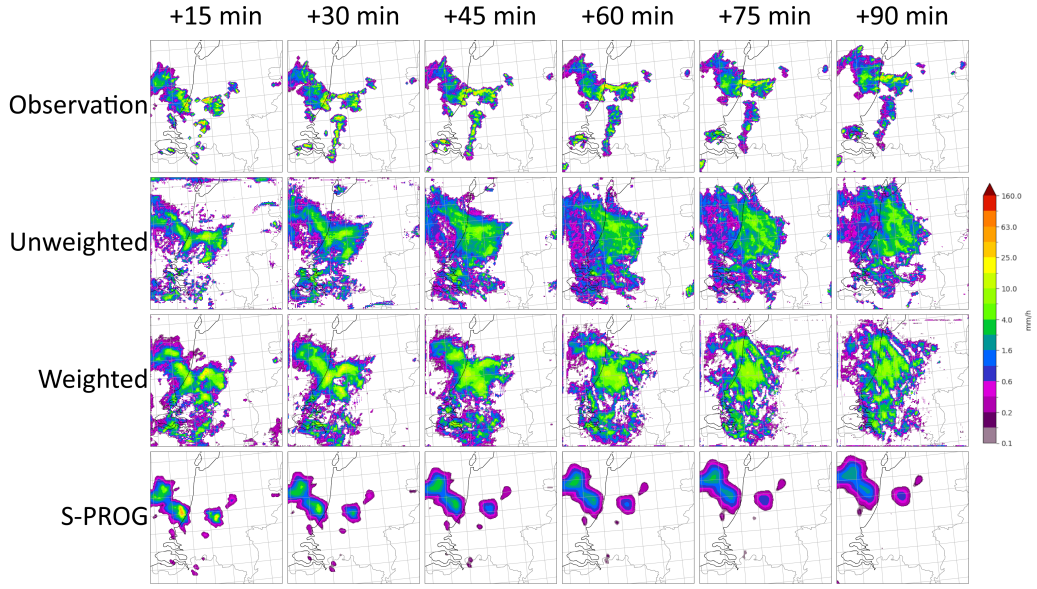

Figure 5.15.: Observation and unweighted and weighted DGMR and S-PROG predictions for Event 4, at t_0 2022-08-17 01:55 UTC.

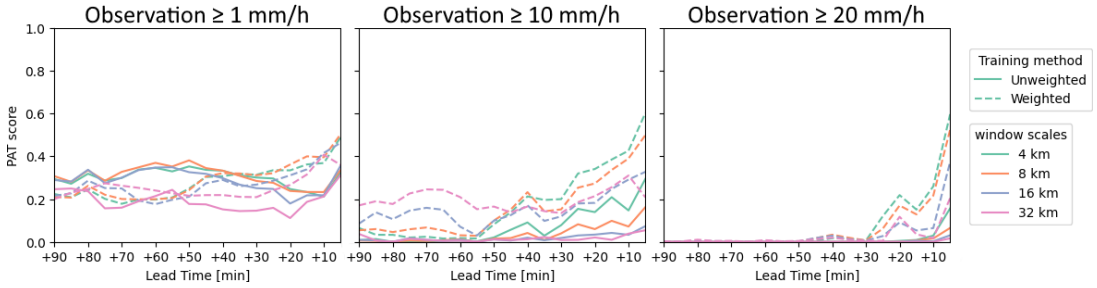

Figure 5.16.: The PAT-score for Event 4 of an ensemble of 10 of each model, at t_{+90}.

Forecast Event 4

In Figure 5.15, S-Prog forecasts rapidly damping intensities, rendering it ineffective for this convective event. Both DGMR models exhibit a similar structure and capture high-intensity cells up to t_{+30}, despite the overall growth in precipitation field size in the forecasts. The weighted model performs best for the highest intensities but transitions to a parallel line structure from t_{+75} onward, forecasting Eastward motion despite the observed Northward movement of the precipitation field. Figure 5.16's PAT-score differs in this case, with smaller window scales often yielding higher scores due to the small size of convective cells and their intense precipitation. The weighted model occasionally predicts these peak intensities at short lead times and on small window scales, but struggles with larger window scales, where predicting peak intensities with a factor of 1.5 remains challenging. Additionally, both models experience intensity dampening at longer lead times, hindering their ability to capture peak intensities of convective cells effectively.

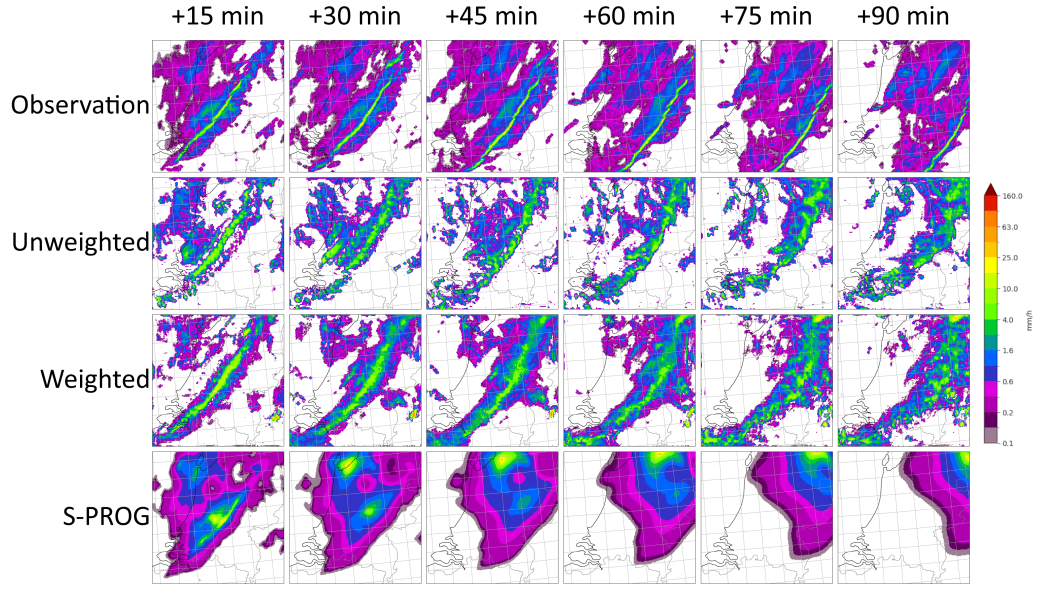

Figure 5.17.: Observation and unweighted and weighted DGMR and S-PROG predictions for Event 5, at t_0 2022-02-20 20:55 UTC.

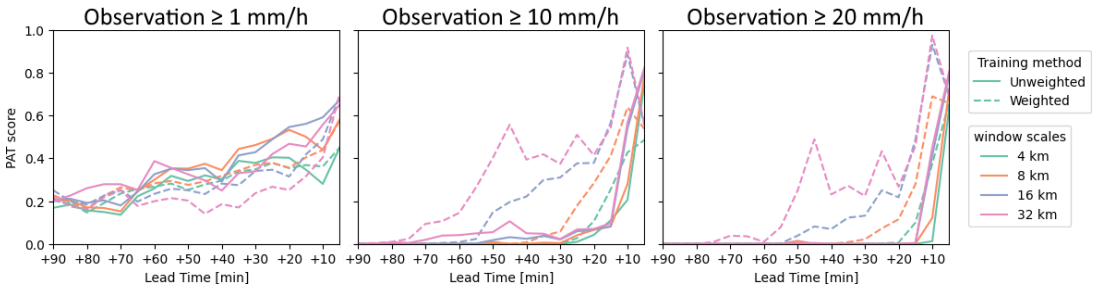

Figure 5.18.: The PAT-score for Event 5 of an ensemble of 10 of each model, at t_{+90}.

Forecast Event 5

In Figure 5.17, S-PROG's forecast for the cold front is very poor, failing to capture the line with high-intensity precipitation and placing it elsewhere within the precipitation field at longer lead times. Both DGMR models initially predict higher precipitation rates near the cold front until t_{+60}, after which they forecast the line breaking up into multiple high-intensity parts instead of remaining coherent. Additionally, the motion slows down, causing both models to fall short of the observed position of the cold front to the East. Figure 5.18's PAT-score reflects this performance, with both models performing poorly at long lead times when observations are $\geq$ 10 mm/h and $\geq$ 20 mm/h. The weighted model shows improvement from t_{+75} with a window scale of 32 km and from t_{+55} with a window scale of 16 km, attributed to the slower motion of the forecast. Both models achieve higher scores at shorter lead times, with the weighted model reaching a PAT-score over 0.9 at t_{+10} for $\geq$ 10 mm/h and $\geq$ 20 mm/h, and up to 0.8 at 8 km and 0.4 at 4 km window scales.This indicates the weighted model's ability to accurately forecast the peak intensities of the cold front at short lead times, both in intensity and location.

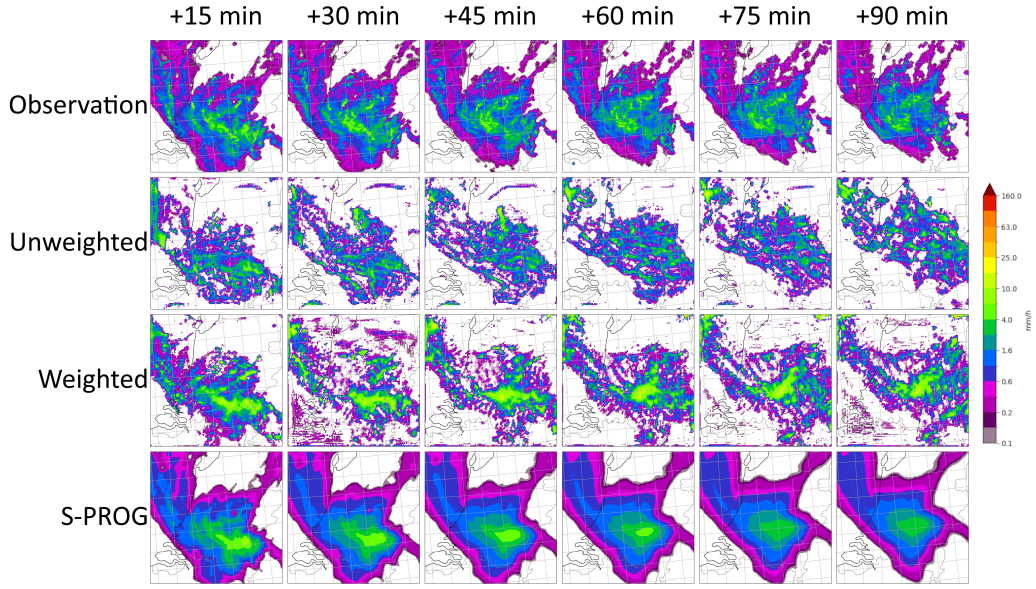

Figure 5.19.: Observation and unweighted and weighted DGMR and S-PROG predictions for Event 6, at t_0 2022-06-05 15:00 UTC.

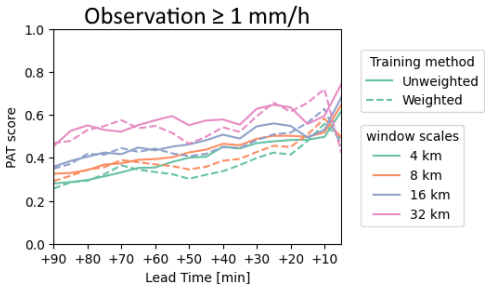

Figure 5.20.: The PAT-score for Event 6 of an ensemble of 10 of each model, at t_{+90} where there is no observed precipitation intensity of $\geq$ 10 mm/h.

Forecast Event 6

In Figure 5.19, S-PROG produces the overall best forecast regarding the outline and movement of the precipitation field. Both DGMR models fail to accurately forecast the rotation and Northward movement of the precipitation field, breaking it up instead. Since the observation at t_{+90} has no precipitation intensity over 10 mm/h, only the PAT-score of $\geq$ 1 mm/h could be determined, shown in Figure 5.20. Scores for both models improve slightly at shorter lead times, with the unweighted model performing slightly better on average across different window scales. However, at t_{+5}, the PAT-score of the weighted model drops due to overestimation, which fades with lead time, dropping within a factor of 1.5 of the observed precipitation rate. In contrast, the unweighted model score jumps as it predicts within a factor of 1.5 of the observed precipitation rates correctly at this lead time.

6

Discussion

Here the results from this research are further discussed. In Section 6.1 the results of the speckle-like clutter removal strategy are discussed as well as the importance within operational nowcasting algorithms. The use of performance metrics in previous work is discussed in Section 6.2. Furthermore the results of this research are compared to the results from others. The splitting of the dataset as well as the comparison of the two training strategies for the DGMR to S-PROG are further discussed in Section 6.3. Finally, the recommendations on further research related to nowcasting are given in Section 6.4

6.1. Clutter removal

As stated in Section 3.1 the real time radar reflectivity product contains many error sources, some with high reflectivity values that would give high precipitation rates. Dekker (2022) has shown that training a nowcasting model on precipitation rates while the data contains clutter could result in the model trying to predict the clutter. From that research it was concluded that training on precipitation rates could only work if more extensive clutter removal was performed. Several clutter removal strategies on radar images with precipitation rates have been suggested by Schreurs (2021). They ended up discarding images from the dataset based on the gradients of the pixel values. This would remove images containing clutter from the training dataset. However the removal of input images to remove clutter cannot be done on operational nowcasting systems.

The morphological clutter removal scheme, proposed in Section 4.1 to remove speckle-like clutter, could be performed within an operational nowcasting system. When this method was applied to the KNMI's reflectivity dataset, from Section 5.1.1 it could be concluded that it would remove some of the errors caused by ship tracks. However errors caused by ground clutter, such as wind farms, are not removed for all settings that were tried. This indicates that the remaining clutter is not speckle-like, but occurs either as large clutter fields, in the case of with the wind farms, or within precipitation fields. Other clutter removal methods are required to improve the dataset by removing these occurrences of clutter.

When looking at what precipitation intensities are removed from radar images in Section 5.1.2, it was shown that more erosion steps remove more speckle-like clutter. Parts of precipitation fields or entire small precipitation fields can also be removed. More erosion steps would result in larger parts of precipitation fields and even more small precipitation fields being removed, which would reduce the quality of the data. It also showed that more erosion steps than dilation steps would remove more of the edge of precipitation fields, consisting mostly of low precipitation rates, influencing the shape of the precipitation fields in the cleaned dataset as well as their size. Therefore a balance is required to preserve most of the true precipitation while removing most speckle-like clutter from the dataset. This balance can be found with 3 erosion steps and 3 dilation steps.

6.2. Performance metrics

In this study the DGMR model performance was determined using two continuous metrics, MAE and MSE, and five categorical metrics in the form of CSI, POD, FAR, F1 and FSS. The continuous metrics have been used by others before on the DGMR (Cambier van Nooten et al., 2023; Elsmann, 2023; Frenkiel, 2022; Ravuri et al., 2021) and are in general often used as a metric for nowcasting models. These have also used some of the categorical metrics. However, the dataset it is applied to as well as which threshold values used vary quite a bit (Woo and Wong, 2017; Niu et al., 2021; Jonnalagadda and Hashemi, 2023). Some compute these metrics on the radar reflectivity values in dBZ instead of precipitation rates in mm/h, applying these metrics on a logarithmic scale instead of a linear one. Some use 0.1 mm/h as lowest threshold with these metrics, which indicates the performance of detecting precipitation, such as Elsmann (2023). Others use higher threshold values and don't give an indication of their models performance of detecting precipitation. These threshold values could be 0.5 or 1 mm/h as their lowest threshold, as done by Frenkiel (2022) and Cambier van Nooten et al. (2023) respectively. These different thresholds and scales could lead to different results and should be selected depending on the goal of the study.

Frenkiel (2022) applied the pre-trained DGMR model to forecast precipitation in the Netherlands and evaluated its performance with the FSS. Elsmann (2023) trained the model and also evaluated it with FSS, both with a threshold of 1 mm/h as done with this study. Unfortunately, the scales used differ a bit and only some window scales are the same. However, the overall shape of the different window scales follow each other. Where Frenkiel shows a linear decrease in FSS over lead time, Elsmann and the results from this study show an initial decrease in FSS which then becomes more stable with lead time. It has to be mentioned that the results from Frenkiel indicate more skilful forecasts over all lead times. This is likely due to the decreased model size from this study. Also, the models for both this study as well as Elsmann's were trained on about four times fewer iterations compared to the original model. Furthermore, recently Antonio and Aitchison (2023) showed cases where the FSS would indicate a skilful forecast, when in fact the forecast and observation were negatively correlated. Since the FSS has been determined by taking the average over 1000 randomly selected events, it is possible that some cases where this occurs are included.

The results show a slight improvement for the weighted training strategy compared to the unweighted strategy on the CSI and F1. The POD is significantly improved with the weighted model due to the tendency to overestimate. However, the MAE, MSE and FAR show that the unweighted model has a higher score. For the MAE and MSE this is due to the unweighted model being trained on all event types and intensities about as often as they occur in the dataset. This is also reflected in the unweighted model having a higher score at

longer lead times according to the FSS with a threshold of 1 mm/h, whereas the weighted model has higher scores for higher threshold values.

The PAT-score indicates an improved ability to forecasting the maximum observed precipitation rate over 10 mm/h within a factor of 1.5 more accurately with the weighted model, even for long lead times in some events. However, for precipitation rates over 1 mm/h, there is only a small improvement for short lead times on some events. In the few test events where precipitation rates of more than 20 mm/h occurred on the frame tested, both models show low forecasting skill at long lead times. The weighted model shows, at a longer lead time than the unweighted model, skill for forecasting these maximum precipitation rates at large window scales. Only at very short lead times does the unweighted model give some skill for these high-intensity precipitation rates. With small window scales, both models only show some skill at very short lead times. There is however one exception, in the case of a convective event with many small precipitation fields with precipitation rates over 10 mm/h the weighted model has more skill with smaller window scales compared to larger window scales. This is due to the model having difficulty on long lead times with forecasting these high intensities and thus only the small window scales being correct when the precipitation rates are lower. Indicating a forecast as correct when it is within a factor of the observed intensity could be a better indication of forecasting skill compared to metrics where the forecast has to be higher than a given threshold value. In the latter case, a forecast of 100 mm/h could be considered correct even if the observation was 10 mm/h, provided the threshold was set lower than or equal to the observed precipitation rate. Furthermore, the PAT-score also indicates when the model starts to become skilful in lead time, similar to FSS.

6.3. Event selection and model training

In some earlier work on the DGMR model, done by Ravuri et al. (2021) and Elsmann (2023), the data was split differently. They selected a whole year for the test set and the remaining years are split over the training and validation set by selecting the first day of each month for the validation set and the other days in the training set. In other research using the DGMR model, by Cambier van Nooten et al. (2023), the dataset was split similarly as with this research, one year for the test set, one year for the validation set and remaining years for the training set. Radar frames close in time to each other are similar, sharing a lot of information that the frames contain. Splitting the training and validation data such that they often are close to each other in time, would lead to using the information of the validation set in the training process, resulting in data leakage (Liu et al., 2022). This would give the model information about future samples and thus an improved performance on the validation set compared to a real-world scenario where it does not have the information about future samples.

Others have trained GAN's for nowcasting without the use of weights during the training process as the focus often lies on improving the model on average precipitation events (Jing et al., 2019; Choi and Kim, 2022; Choi et al., 2023). These are often of smaller scale with lower maximum precipitation intensities. Selecting events with precipitation based on the ISW, as described in Section 4.2.1, is required to avoid training mainly on dry events which do not have any information for the model to learn from. To then also use these weights during training for selecting the events to train on with each step, allows the training data to be re-balanced towards more extreme events. The results from Section 5.2 have shown that the weighted training strategy improves many metrics for higher precipitation rates, such as the CSI, POD and PAT-score. At the same time, it can also be concluded that this

weighted training strategy results in lower scores on the metrics that indicate the average performance, in the form of the MAE and MSE, showing that there is a trade-off between performance on average events as compared to extreme events.

When comparing the trained models visually to S-PROG, they often under perform with regards to forecasting the motion of the precipitation fields. Where S-PROG smooths and fades with lead time, DGMR has no smoothing and less fading but at the cost of introducing parallel line structures when the observed motion differs a lot from the average wind direction. The comparison is not completely fair, as the DGMR models only make forecasts using information within the defined model domain of 256×256 whereas S-PROG constructs the forecast using the image extent of 700×765 and only being evaluated on the smaller domain. This leads to S-PROG having more information to make forecasts for longer lead times near the edge of the domain, therefor the focus of this comparison should be on the research domain to reduce this effect.

To make forecasts over larger areas would require the model to be trained on a mosaic, generalizing the model to the entire area. Another approach is by having multiple models, each trained on a different partly overlapping section and learning precipitation patterns caused by local conditions which are then combined into a single forecast over a larger area. Furthermore, the reduced DGMR model size used in this research has a significant impact on the performance, for instance forecasting the motion at long lead times to always be towards the East. Zou (2023) looked into the effect of a reduced model size for a machine learning nowcasting model and concluded that the reduced models were able to reproduce similar structures to the original model size. There was however no mention of the model motion being forecasted towards the same direction at long lead times for all events.

6.4. Recommendations

The dataset used in this research contains errors. Since GAN models try to recreate the structure of the data, it is important to remove these errors. The speckle-like clutter removal should be improved further, and additional data cleanup methods to remove other types of errors should be investigated. Other types of errors that still pose issues in the current dataset include radar spikes, large clutter fields, and clutter on fixed locations that occur within precipitation fields.

Due to the higher occurrence of errors over the sea, events with the highest weight contained many errors when the research domain shown in Figure 4.5 only had a 32 km boundary between it and the model input. Consequently, it was reduced in size to ensure that errors over the sea, which have not been fully removed, did not influence the event weights and the scores calculated when verifying the model performance. Increasing the size of the research domain by following the coastline on the west side could expand the area on which verification is done. This could potentially providing a better indication of the model performance.

To sample more complex high precipitation rate events during training, the event weighting could be further improved. The models currently have issues with forecasting motion, which could be addressed within the weighting calculation by giving a higher weight to events with non-homogeneous motion, such as rotation.

Other ways of making the training-validation-test split should be investigated. From Figure 4.9 it can be seen that the distribution of the top 20% events does not follow the same

distribution for each subset. One way to improve this while still preventing data leakage is to add all consecutive top 20% events to one subset at a time. When the weight has dropped lower for a long enough duration, the next consecutive top 20% events can then be added to another subset, reducing data leakage while spreading the data more evenly among the subsets. Since the data preprocessing performed by KNMI as well as the changed radar location, the current split results in the validation and test sets not containing any data from the old setup while the training data has 9 years of the old setup and only 2 full years of the new setup. This would be resolved with the proposed method constructing the training-validation-test split.

During training, the model is currently penalized by the discriminators over the entire model input. Over longer lead times this causes issues near the edges where the model lacks context outside of the model input to make predictions. This may introduce artifacts into the predictions near the edges. A possible method to prevent this is to decrease the area on which the discriminators penalize the generator as lead time increases.

The model is currently limited to a fixed domain of 256×256 km. To make forecasts over a larger area would require a different approach. Multiple models could be trained on different, partly overlapping sections to learn local precipitation patterns and have their forecasts combined into one forecast for the larger area. Another approach is to train a single model on the different sections, generalizing the model forecasts over the entire area.

The generator and discriminators are reduced in size due to limited computational resources. The quality of the forecasts can likely be improved by increasing the model to its original size. Another improvement in forecasting high-intensity precipitation rates could be achieved by providing more context, as was done by Elsmann (2023) with the inclusion of ETH.

Many metrics commonly used in evaluating nowcasting models indicate the average performance, but few exist that evaluate the performance on high-intensity precipitation specifically, which is important for early warning systems. More evaluation methods are needed that can evaluate the performance for different goals, such as predicting high intensities or total hourly precipitation in a catchment. The metrics most often used, such as the MAE, MSE, and CSI are pixel-wise comparisons that may not be a fair evaluation depending on the goal of the user. This should be taken into consideration when picking evaluation metrics.

The two training sampling strategies applied in this research show similar performance on these common metrics, where they both have low scores. The only metric with a clear difference between the two models is the PAT-score, which was specifically adapted to show how often and with which lead time the peak precipitation rate could be forecasted correctly. Since other metrics show similar performance, further investigation of the training sampling is required to check if it improves the forecast of high-intensity precipitation rates.

The DGMR is also able to forecast an ensemble due to the inclusion of the Latent Conditioning Stack in the model, introducing Gaussian noise when generating forecasts. The quality and distribution of this ensemble should be investigated. How similar the ensemble members are and if an ensemble captures the observation within its envelope for all lead times is not yet known. Further investigation may lead to further improvements in ensemble nowcasting for early warning systems.

7

Conclusion

This research aimed to improve the precipitation nowcast of high-intensity precipitation events in the Netherlands using the Deep Generative Model of Radar (DGMR), a generative adversarial network based precipitation nowcasting algorithm. The following questions were addressed:

How can the data be processed to improve the data quality for high-intensity precipitation events in the Netherlands?
The dataset used spanned from 2008 to 2022 and was provided by the KNMI. It includes a composite of radar reflectivity from two radar systems and required considerable preprocessing due to the relocation of one radar system and updates to the processing done by the KNMI. A speckle-like clutter removal method was applied to remove most speckle-like clutter, which reduced errors caused by ship tracks and ground clutter over land. This method improved the data quality of high-intensity precipitation rates caused by erroneous sources, albeit at the cost of removing real precipitation fields smaller than 6 km. Several other types of errors in larger structures, such as wind farms, were not removed. Further data preprocessing is required to remove more errors while minimizing the removal of real precipitation fields.

In what ways can a Generative Adversarial Network (GAN) be trained with radar images to improve the prediction for high-intensity precipitation events in the Netherlands?
The DGMR was trained on the cleaned radar images using two training sampling strategies. The weighted strategy prioritizes sampling events with higher precipitation rates, resulting in better performance on Critical Success Index (CSI), Probability of Detection (POD), F1 and Peak Anticipation Time (PAT)-score. The unweighted strategy sampled events uniformly, resulting in better scores on Mean Absolute Error (MAE), Mean Squared Error (MSE), and False Alarm Rate (FAR). These results show that the training sampling strategy influences the model's performance, with the weighted strategy improving the forecast for high-intensity precipitation rates.

How does this model, under different training strategies, compare to S-PROG, a state-of-the-art extrapolation based nowcasting system?
To compare the DGMR models' forecast to those from S-PROG, six precipitation events with different behaviours were selected. These comparisons show that DGMR can forecast

high precipitation rates at longer lead times without blurring. However, S-PROG performed better with forecasting the motion as well as lower precipitation rates for most events. Although S-PROG outperforms DGMR in forecasting motion and low precipitation rates, the DGMR models provide more detail and can forecast higher precipitation rates at longer lead times.

Bibliography

Aberson, K. (2011). The spatial and temporal variability of the vertical dimension of rainstorms and their relation with precipitation intensity. Internal Rep. IR 2011-03. *KNMI*.

Adabi, M., Barhab, P., Chen, J., Chen, Z., Davis, A., Dean, J., Zheng, X., et al. (2016). Tensorflow: A System for Large-Scale Machine Learning (vol. 16, pp. 265–283). In *12th USENIX Symposium on Operating Systems Design and Implementation, USENIX Association*.

Agrawal, S., Barrington, L., Bromberg, C., Burge, J., Gazen, C., and Hickey, J. (2019). Machine Learning for Precipitation Nowcasting from Radar Images. *arXiv:1912.12132*.

Antonio, B. and Aitchison, L. (2023). The Fractions Skill Score doesn't always measure skill. *arXiv:2311.11985*.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein GAN. *arXiv:1701.07875*.

Ayzel, G., Heistermann, M., Sorokin, A., Nikitin, O., and Lukyanova, O. (2019). All convolutional neural networks for radar-based precipitation nowcasting. *Procedia Computer Science*, 150:186–192.

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.

Beekhuis, H. and Holleman, I. (2008). From Pulse to Product, Highlights of the digital-if upgrade of the Dutch national radar network. ERAD 5, Helsinki, KNMI.

Berenguer, M., Surcel, M., Zawadzki, I., Xue, M., and Kong, F. (2012). The Diurnal Cycle of Precipitation from Continental Radar Mosaics and Numerical Weather Prediction Models. Part II: Intercomparison among Numerical Models and with Nowcasting. *Monthly weather review*, 140(8):2689–2705.

Bi, H. (2022). Extreme Precipitation Nowcasting using Deep Generative Models. Master's thesis, Delft University of Technology.

Bihlo, A. (2020). A generative adversarial network approach to (ensemble) weather prediction. *arXiv:2006.07718*.

Brock, A., Donahue, J., and Simonyan, K. (2018). Large Scale GAN Training for High Fidelity Natural Image Synthesis. *arXiv:1809.11096*.

Cambier van Nooten, C., Schreurs, K., Wijnands, J. S., Leijnse, H., Schmeits, M., Whan, K., and Shapovalova, Y. (2023). Improving precipitation nowcasting for high-intensity events using deep generative models with balanced loss and temperature data: a case study in the Netherlands. *Artificial Intelligence for the Earth Systems*.

Choi, J., Kim, Y., Kim, K., Jung, S., and Cho, I. (2023). PCT-CycleGAN: Paired Complementary Temporal Cycle-Consistent Adversarial Networks for Radar-Based Precipitation Nowcasting. *arXiv:2211.15046*.

Choi, S. and Kim, Y. (2022). Rad-cGAN v1. 0: Radar-based precipitation nowcasting model with conditional generative adversarial networks for multiple dam domains. *Geoscientific Model Development*, 15(15):5967–5985.

Clark, A., Donahue, J., and Simonyan, K. (2019). Adversarial Video Generation on Complex Datasets. *arXiv:1907.06571*.

Dekker, D. (2022). Perceptual losses in precipitation nowcasting: Exploring limits and potential. Master's thesis, Delft University of Technology.

Delft High Performance Computing Centre (DHPC) (2024). DelftBlue Supercomputer (Phase 2). https://www.tudelft.nl/dhpc/system.

Douris, J., Alexeeva, V., Shaw, B., and Ikeda, R. (2023). WMO Atlas of Mortality and Economic Losses from Weather, Climate, and Water Extremes (1970–2019). *Weather Meteorological Organization*, (1267).

Doviak, R. J. and Zrnić, D. S. (1993). *Doppler radar and weather observations*. Academic Press, second edition.

Duncan, J., Subramanian, S., and Harrington, P. (2022). Generative Modeling of High-resolution Global Precipitation Forecasts. *arXiv:2210.12504*.

Elsmann, F. (2023). Precipitation Nowcasting: Exploring the Impact of Echo Top Heights in Generative Models. Master's thesis, Radboud University.

Foresti, L., Reyniers, M., Seed, A., and Delobbe, L. (2016). Development and verification of a real-time stochastic precipitation nowcasting system for urban hydrology in Belgium. *Hydrology and Earth System Sciences*, 20:505–527.

Frenkiel, Y. (2022). Validating a Deep Generative Precipitation Nowcasting Model on the Netherlands. Master's thesis, Utrecht University.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Nets. *Advances in neural information processing systems*, 27.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2020). Generative Adversarial Networks. *Communications of the ACM*, 63(11):139–144.

Grecu, M. and Krajewski, W. (2000). A large-sample investigation of statistical procedures for radar-based short-term quantitative precipitation forecasting. *Journal of Hydrology*, 239(1):69–84.

Han, D., Choo, M., Im, J., Shin, Y., Lee, J., and Jung, S. (2023). Precipitation nowcasting using ground radar data and simpler yet better video prediction deep learning. *GIScience & Remote Sensing*, 60(1):2203363.

Hayatbini, N., Kong, B., Hsu, K., Nguyen, P., Sorooshian, S., Stephens, G., Fowlkes, C., Nemani, R., and Ganguly, S. (2019). Conditional generative adversarial networks (cGANs) for near real-time precipitation estimation from multispectral GOES-16 satellite imageries—PERSIANN-cGAN. *Remote Sensing*, 11(19):2193.

Holleman, I. and Beekhuis, H. (2005). Review of the KNMI clutter removal scheme. Technical Report TR-284, KNMI.

Imhoff, R., Brauer, C., Overeem, A., Weerts, A., and Uijlenhoet, R. (2020). Spatial and Temporal Evaluation of Radar Rainfall Nowcasting Techniques on 1,533 Events. *Water Resources Research*, 56(8):e2019WR026723.

Imhoff, R. O., Brauer, C. C., van Heeringen, K., Uijlenhoet, R., and Weerts, A. H. (2022). Large-Sample Evaluation of Radar Rainfall Nowcasting for Flood Early Warning. *Water Resources Research*, 58(3):e2021WR031591.

Isola, P., Zhu, J., Zhou, T., and Efros, A. A. (2018). Image-to-Image Translation with Conditional Adversarial Networks. *arXiv:1611.07004v3*.

Ji, Y., Gong, B., Langguth, M., Mozaffari, A., and Zhi, X. (2023). CLGAN: a generative adversarial network (GAN)-based video prediction model for precipitation nowcasting. *Geoscientific Model Development*, 16(10):2737–2752.

Jing, J., Li, Q., Ding, X., Sun, N., Tang, R., and Cai, Y. (2019). AENN: A Generative Adversarial Neural Network for Weather Radar Echo Extrapolation. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42:89–94.

Jonnalagadda, J. and Hashemi, M. (2023). Quality-Aware Conditional Generative Adversarial Networks for Precipitation Nowcasting. *Engineering Proceedings*, 39(1):11.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2017). Progressive Growing of GANs for Improved Quality, Stability, and Variation. *arXiv:1710.10196v3*.

Lebedev, V., Ivashkin, V., Rudenko, I., Ganshin, A., Molchanov, A., Ovcharenko, S., Grokhovetskiy, R., Bushmarinov, I., and Solomentsev, D. (2019). Precipitation Nowcasting with Satellite Imagery. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2680–2688.

Leijnse, H., Beekhuis, H., and Holleman, I. (2016). Doppler clutter removal on KNMI weather radars. Technical Report TR-355, KNMI.

Lenderink, G. and Van Meijgaard, E. (2010). Linking increases in hourly precipitation extremes to atmospheric temperature and moisture changes. *Environmental Research Letters*, 5(2):025208.

Lin, G. (2022). Nowcasting of Extreme Rainfall in Dutch Cities. Master's thesis, Delft University of Technology.

Liu, F., Chen, L., Zheng, Y., and Feng, Y. (2022). A Prediction Method with Data Leakage Suppression for Time Series. *Electronics*, 11(22):3701.

Liu, H. and Lee, I. (2020). MPL-GAN: Toward realistic meteorological predictive learning using conditional GAN. *IEEE Access*, 8:93179–93186.

Manola, I., Steeneveld, G., Uijlenhoet, R., and Holtslag, A. A. (2020). Analysis of urban rainfall from hourly to seasonal scales using high-resolution radar observations in the Netherlands. *International Journal of Climatology*, 40(2):822–840.

Marshall, J. S. and Palmer, W. M. K. (1948). The distribution of raindrops with size. *Journal of Atmospheric Sciences*, 5(4):165–166.

Mirza, M. and Osindero, S. (2014). Conditional Generative Adversarial Nets. *CoRR*, abs/1411.1784.

Niu, D., Huang, J., Zang, Z., Xu, L., Che, H., and Tang, Y. (2021). Two-Stage Spatiotemporal Context Refinement Network for Precipitation Nowcasting. *Remote Sensing*, 13(21):4285.

Overeem, A., de Vries, H., Sakka, H. A., Uijlenhoet, R., and Leijnse, H. (2021). Rainfall-Induced Attenuation Correction for Two Operational Dual-Polarization C-Band Radars in the Netherlands. *Journal of Atmospheric and Oceanic Technology*, 38(6):1125 – 1142.

Overeem, A., Uijlenhoet, R., and Leijnse, H. (2020). Full-Year Evaluation of Nonmeteorological Echo Removal with Dual-Polarization Fuzzy Logic for Two C-Band Radars in a Temperate Climate. *Journal of Atmospheric and Oceanic Technology*, 37(9):1643–1660.

Pierce, C., Seed, A., Ballard, S., Simonin, D., and Li, Z. (2012). Nowcasting. In *Doppler Radar Observations-Weather Radar, Wind Profiler, Ionospheric Radar, and Other Advanced Applications*. IntechOpen.

Pulkkinen, S., Nerini, D., Pérez Hortal, A. A., Velasco-Forero, C., Seed, A., Germann, U., and Foresti, L. (2019). Pysteps: an open-source Python library for probabilistic precipitation nowcasting (v1.0). *Geoscientific Model Development*, 12(10):4185–4219.

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., et al. (2021). Skillful Precipitation Nowcasting using Deep Generative Models of Radar. *Nature*, 597:672–677.

Roberts, N. M. and Lean, H. W. (2008). Scale-Selective Verification of Rainfall Accumulations from High-Resolution Forecasts of Convective Events. *Monthly Weather Review*, 136(1):78 – 97.

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.

Saito, M., Saito, S., Koyama, M., and Kobayashi, S. (2018). TGANv2: Efficient Training of Large Models for Video Generation with Multiple Subsampling Layers. *arXiv:1811.09245*.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved Techniques for Training GANs. *arXiv:1606.03498*.

Schreurs, K. (2021). Precipitation Nowcasting using Generative Adversarial Networks. Master's thesis, Radboud University.

Seed, A. (2003). A Dynamic and Spatial Scaling Approach to Advection Forecasting. *Journal of Applied Meteorology and Climatology*, 42(3):381–388.

Sønderby, C. K., Espeholt, L., Heek, J., Dehghani, M., Oliver, A., Salimans, T., Agrawal, S., Hickey, J., and Kalchbrenner, N. (2020). MetNet: A Neural Weather Model for Precipitation Forecasting. *arXiv:2003.12140*.

Van der Kooij, E. (2021). Nowcasting Heavy Precipitation in the Netherlands: a Deep Learning Approach. Master's thesis, Delft University of Technology.

Wang, C., Wang, P., Wang, P., Xue, B., and Wang, D. (2021). Using Conditional Generative Adversarial 3-D Convolutional Neural Network for Precise Radar Extrapolation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:5735–5749.

Wessels, H. R. (2006). KNMI Radar Methods. Technical Report TR-293, KNMI.

Woo, W. and Wong, W. (2017). Operational Application of Optical Flow Techniques to Radar-Based Rainfall Nowcasting. *Atmosphere*, 8(3):48.

Xu, L., Niu, D., Zhang, T., Chen, P., Chen, X., and Li, Y. (2022). Two-Stage UA-GAN for Precipitation Nowcasting. *Remote Sensing*, 14(23):5948.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D. (2016). Stack-GAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. *arXiv:1612.03242*.

Zhang, Y., Long, M., Chen, K., Xing, L., Jin, R., Jordan, M. I., and Wang, J. (2023). Skilful nowcasting of extreme precipitation with NowcastNet. *Nature*, 619:526–532.

Zou, Y. (2023). Diverse Explorations of Rainfall Nowcasting with TrajGRU: Mitigating Smoothness and Fading Out Challenges for Longer Lead Times. Master's thesis, Delft University of Technology.

A

Clutter removal scheme

Some more example radar frames with the clutter removal scheme.

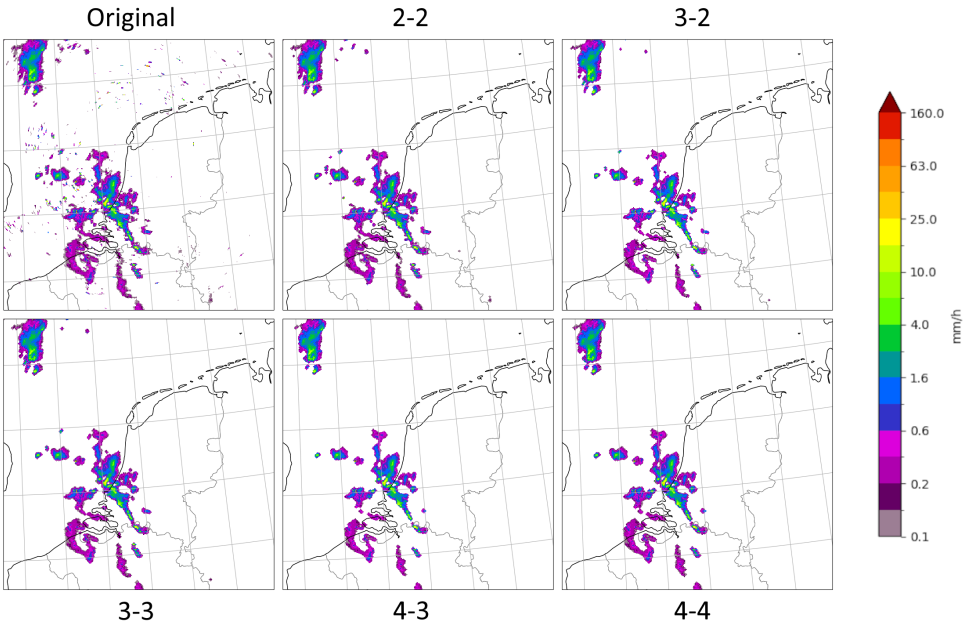

Figure A.1.: Example of a radar image before and after cleanup under different settings taken on June 28th, 2011 at 09:00 UTC with several precipitation fields of different sizes and some speckle like clutter.

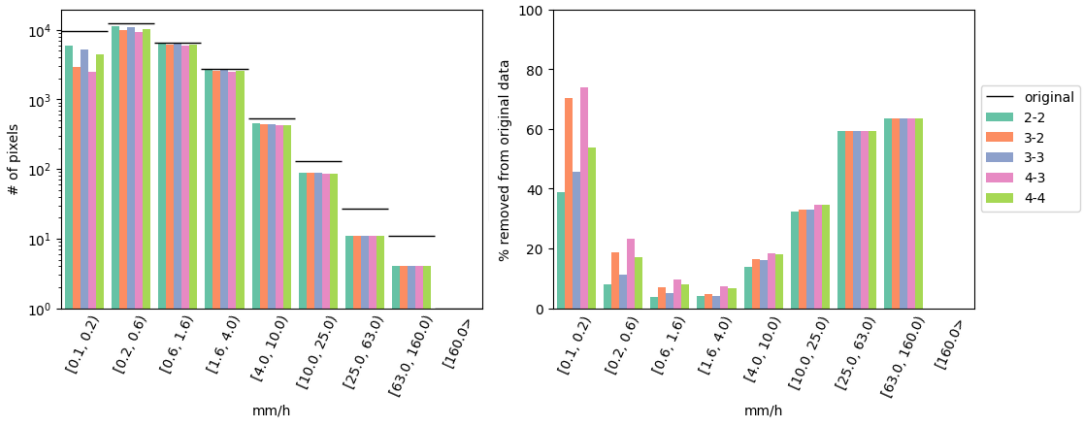

Figure A.2.: Extra detail for Figure A.1. The number of pixels per intensity bin after cleaning with the different settings compared to the original number of pixels (left). The number of pixels per intensity bin after cleaning with the different settings compared to the original number of pixels (right).

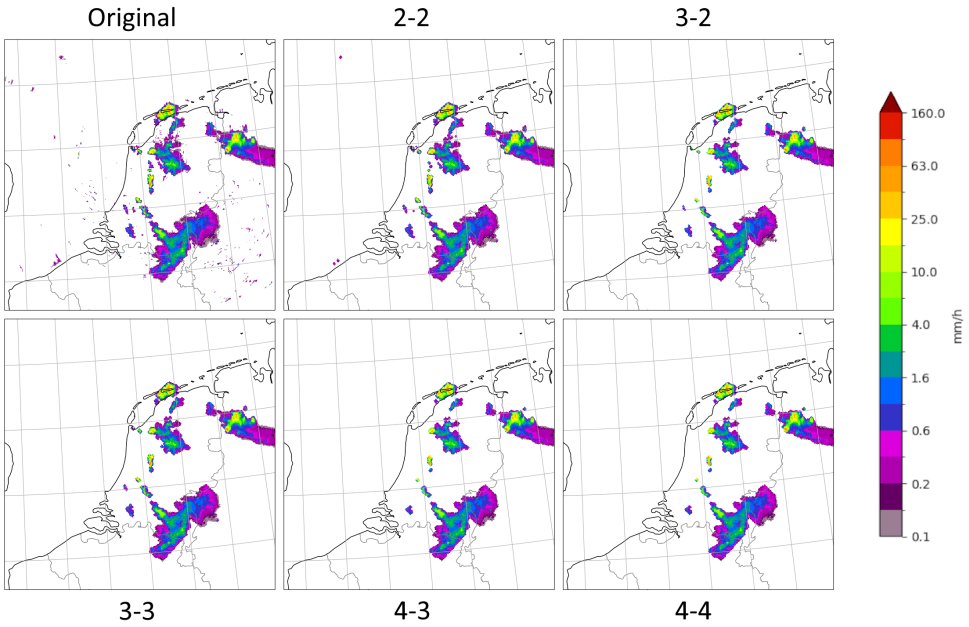

Figure A.3.: Example of a radar image before and after cleanup under different settings taken on May 23rd, 2012 at 20:00 UTC with some small precipitation fields with high intensities and some speckle like clutter.

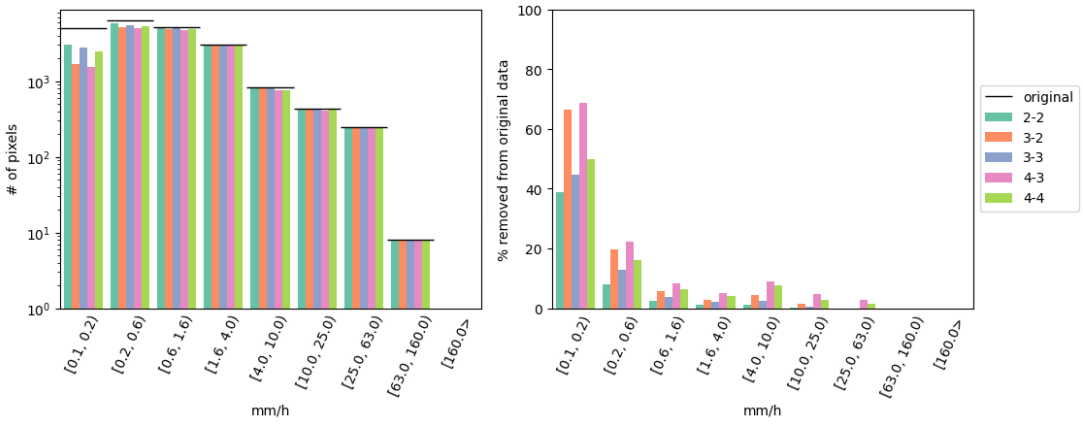

Figure A.4.: Extra detail for Figure A.3. The number of pixels per intensity bin after cleaning with the different settings compared to the original number of pixels (left). The number of pixels per intensity bin after cleaning with the different settings compared to the original number of pixels (right).

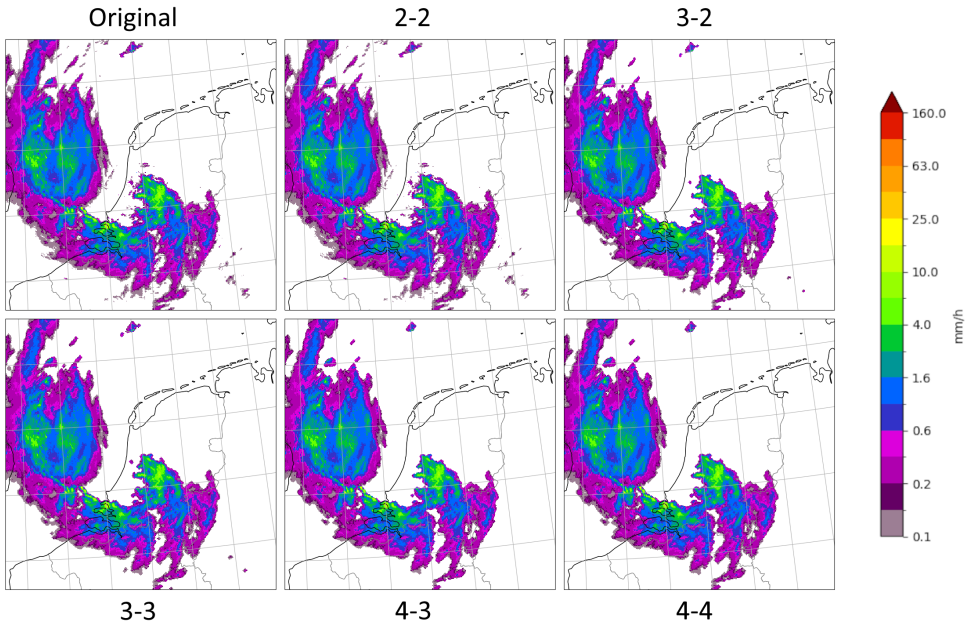

Figure A.5.: Example of a radar image before and after cleanup under different settings taken on October 13th, 2013 at 05:00 UTC with a large precipitation field and no clutter.

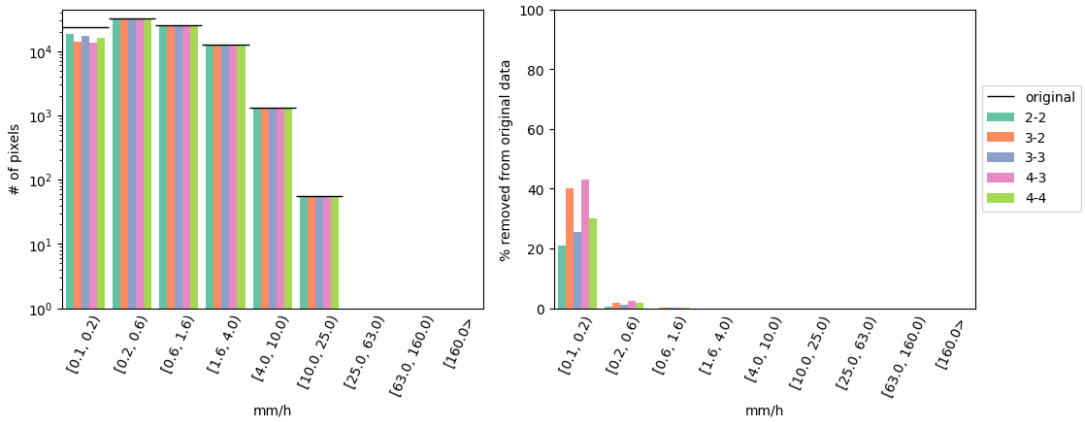

Figure A.6.: Extra detail for Figure A.5. The number of pixels per intensity bin after cleaning with the different settings compared to the original number of pixels (left). The number of pixels per intensity bin after cleaning with the different settings compared to the original number of pixels (right).

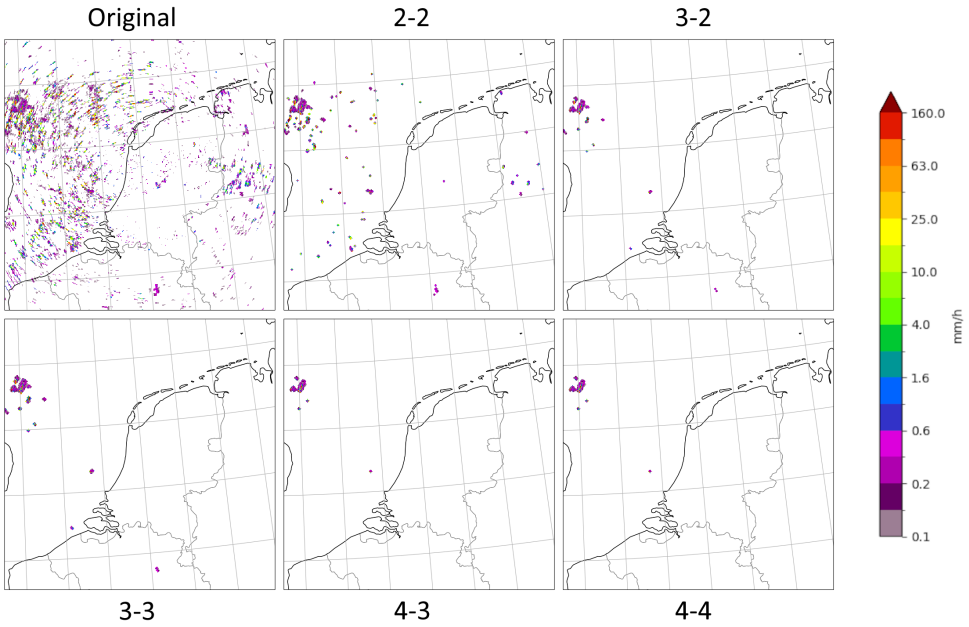

Figure A.7.: Example of a radar image before and after cleanup under different settings taken on March 30th, 2014 at 06:00 UTC with heavy speckle clutter over the entire radar images, most over the ocean.

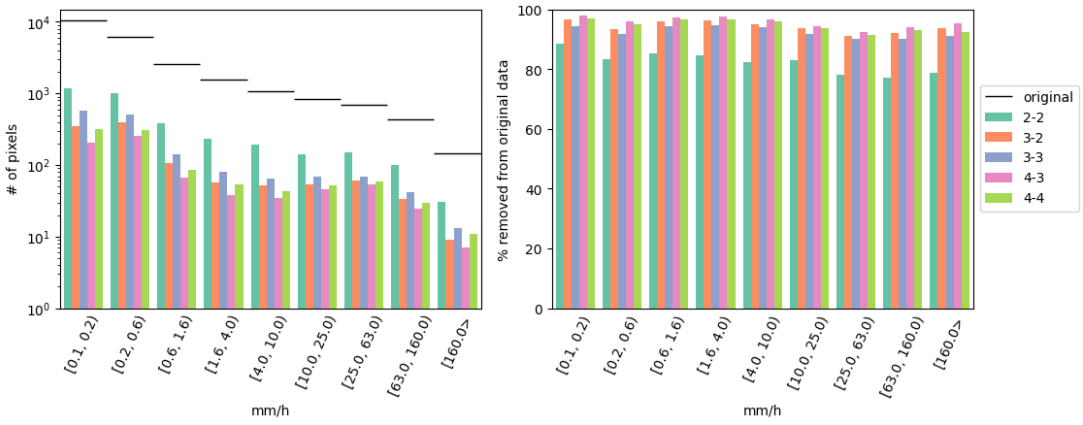

Figure A.8.: Extra detail for Figure A.7. The number of pixels per intensity bin after cleaning with the different settings compared to the original number of pixels (left). The number of pixels per intensity bin after cleaning with the different settings compared to the original number of pixels (right).

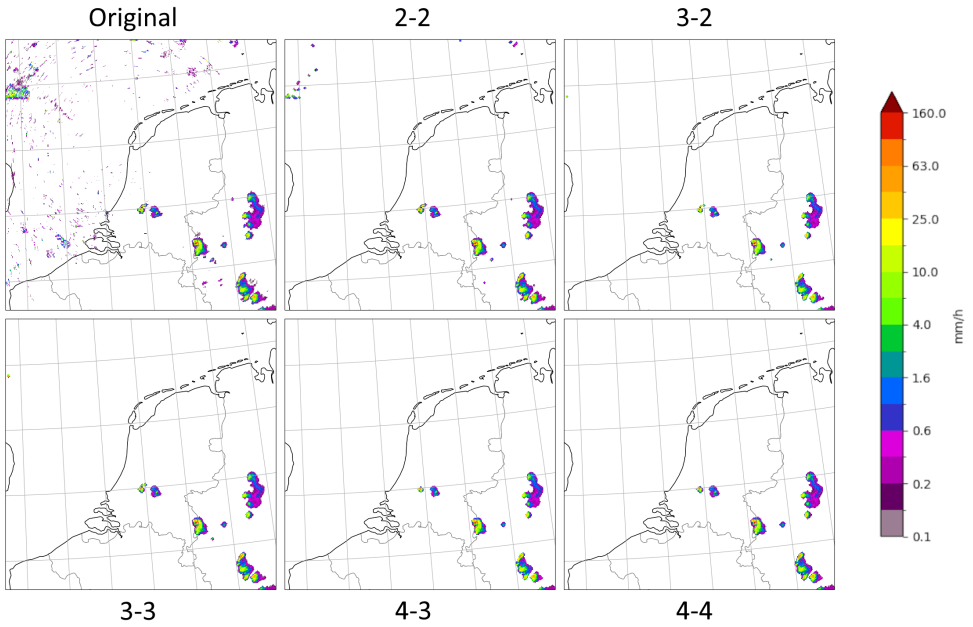

Figure A.9.: Example of a radar image before and after cleanup under different settings taken on August 27th, 2019 at 17:00 UTC with some small precipitation fields with high intensities, speckle like clutter as well as errors caused by wind farms.

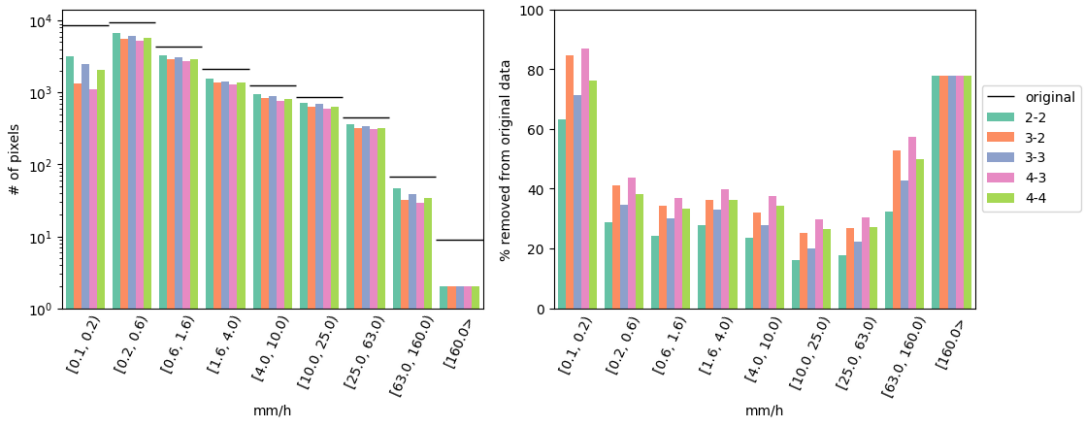

Figure A.10.: Extra detail for Figure A.9. The number of pixels per intensity bin after cleaning with the different settings compared to the original number of pixels (left). The number of pixels per intensity bin after cleaning with the different settings compared to the original number of pixels (right).

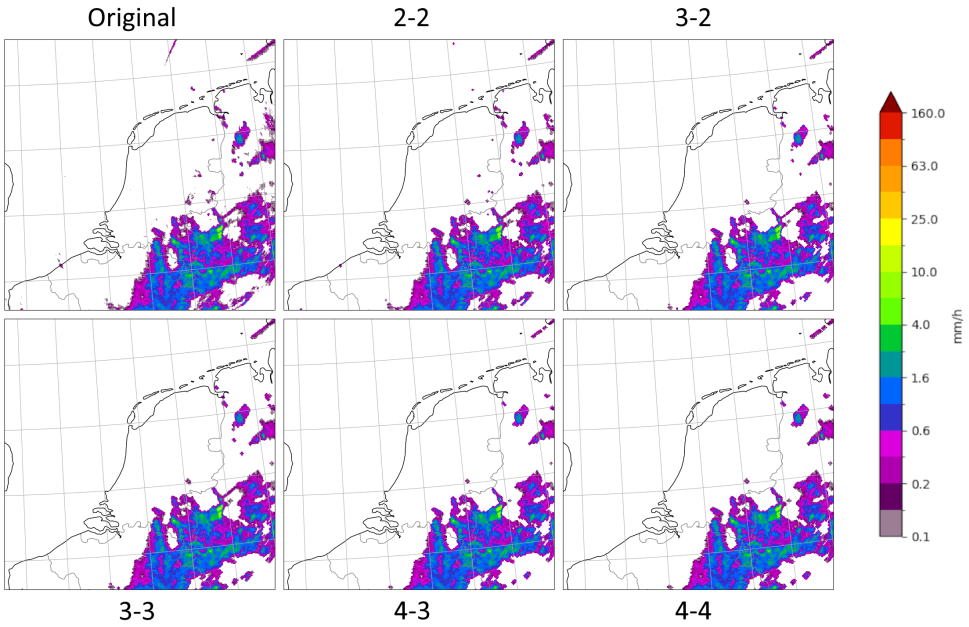

Figure A.11.: Example of a radar image before and after cleanup under different settings taken on July 14th, 2021 at 08:00 UTC with a large precipitation field and a few smaller ones, as well as radar spikes in the top right.

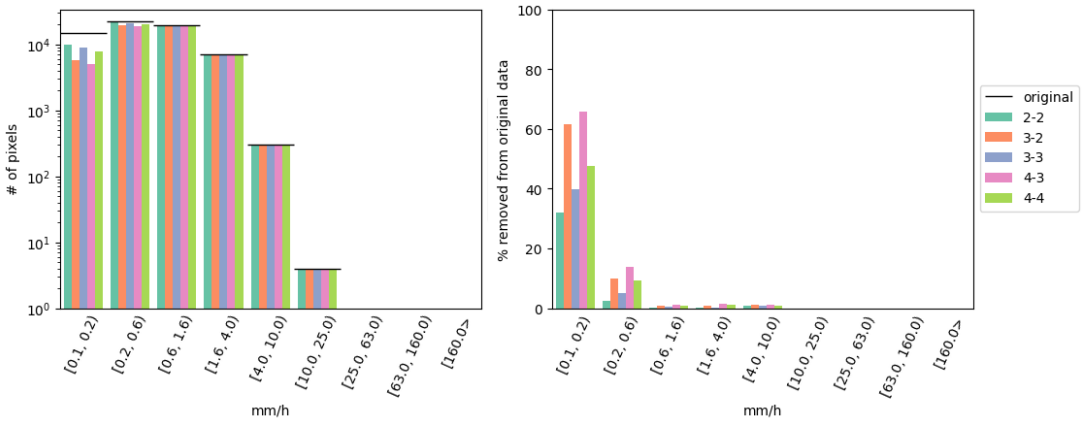

Figure A.12.: Extra detail for Figure A.11. The number of pixels per intensity bin after cleaning with the different settings compared to the original number of pixels (left). The number of pixels per intensity bin after cleaning with the different settings compared to the original number of pixels (right).

B

Metric results split on event weights

Table B.1.: Average categorical metrics on 7 randomly sampled events from the test set that fall in the top 1% of weights for different threshold values on the DGMR model under the two different training strategies. Where (>) indicates higher values are better scores and (<) indicates that lower values are better.

	CSI (>)			POD (>)		
	$\geq$ 1 mm/h	$\geq$ 10 mm/h	$\geq$ 20 mm/h	$\geq$ 1 mm/h	$\geq$ 10 mm/h	$\geq$ 20 mm/h
Unweighted	0.140	0.001	0.000	0.367	0.094	0.000
Weighted	0.076	0.004	0.001	0.551	0.064	0.002
	FAR (<)			F1 (>)		
	$\geq$ 1 mm/h	$\geq$ 10 mm/h	$\geq$ 20 mm/h	$\geq$ 1 mm/h	$\geq$ 10 mm/h	$\geq$ 20 mm/h
Unweighted	0.758	0.999	1.000	0.240	0.002	0.000
Weighted	0.913	0.995	0.999	0.131	0.009	0.001

Table B.2.: Average categorical metrics on 38 randomly sampled events from the test set that fall in the top 2% of weights for different threshold values on the DGMR model under the two different training strategies. Where (>) indicates higher values are better scores and (<) indicates that lower values are better.

	CSI (>)			POD (>)		
	$\geq$ 1 mm/h	$\geq$ 10 mm/h	$\geq$ 20 mm/h	$\geq$ 1 mm/h	$\geq$ 10 mm/h	$\geq$ 20 mm/h
Unweighted	0.135	0.005	0.004	0.445	0.034	0.013
Weighted	0.074	0.006	0.006	0.477	0.042	0.009
	FAR (<)			F1 (>)		
	$\geq$ 1 mm/h	$\geq$ 10 mm/h	$\geq$ 20 mm/h	$\geq$ 1 mm/h	$\geq$ 10 mm/h	$\geq$ 20 mm/h
Unweighted	0.812	0.977	0.984	0.220	0.011	0.008
Weighted	0.910	0.992	0.997	0.131	0.011	0.002

Table B.3.: Average categorical metrics on 177 randomly sampled events from the test set that fall in the top 5% of weights for different threshold values on the DGMR model under the two different training strategies. Where (>) indicates higher values are better scores and (<) indicates that lower values are better.

	CSI (>)			POD (>)		
	$\geq$ 1 mm/h	$\geq$ 10 mm/h	$\geq$ 20 mm/h	$\geq$ 1 mm/h	$\geq$ 10 mm/h	$\geq$ 20 mm/h
Unweighted	0.107	0.009	0.005	0.430	0.042	0.021
Weighted	0.095	0.008	0.004	0.494	0.068	0.038
	FAR (<)			F1 (>)		
	$\geq$ 1 mm/h	$\geq$ 10 mm/h	$\geq$ 20 mm/h	$\geq$ 1 mm/h	$\geq$ 10 mm/h	$\geq$ 20 mm/h
Unweighted	0.855	0.972	0.984	0.178	0.016	0.009
Weighted	0.885	0.988	0.990	0.162	0.016	0.008

Table B.4.: Average categorical metrics on 454 randomly sampled events from the test set that fall in the top 10% of weights for different threshold values on the DGMR model under the two different training strategies. Where (>) indicates higher values are better scores and (<) indicates that lower values are better.

	CSI (>)			POD (>)		
	$\geq$ 1 mm/h	$\geq$ 10 mm/h	$\geq$ 20 mm/h	$\geq$ 1 mm/h	$\geq$ 10 mm/h	$\geq$ 20 mm/h
Unweighted	0.101	0.007	0.003	0.415	0.032	0.014
Weighted	0.099	0.008	0.003	0.505	0.069	0.026
	FAR (<)			F1 (>)		
	$\geq$ 1 mm/h	$\geq$ 10 mm/h	$\geq$ 20 mm/h	$\geq$ 1 mm/h	$\geq$ 10 mm/h	$\geq$ 20 mm/h
Unweighted	0.866	0.972	0.987	0.170	0.013	0.005
Weighted	0.881	0.989	0.993	0.167	0.014	0.006

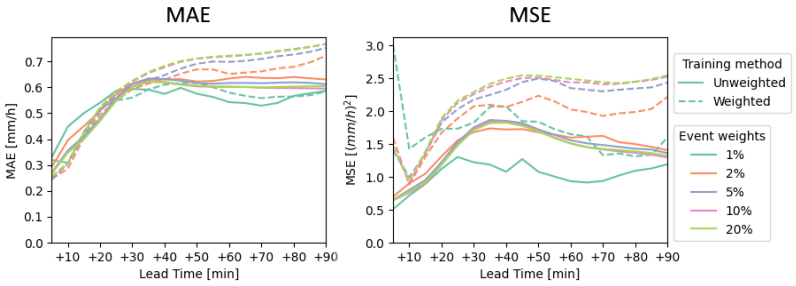

Figure B.1.: Average Mean Absolute Error (MAE) and Mean Squared Error (MSE) on 1000 randomly sampled events, split on event weight thresholds from the test set on the DGMR model under the two different training strategies. 7 top 1% events, 38 top 2% events, 177 top 5% events, 454 top 10% events and 1000 top 20% events.

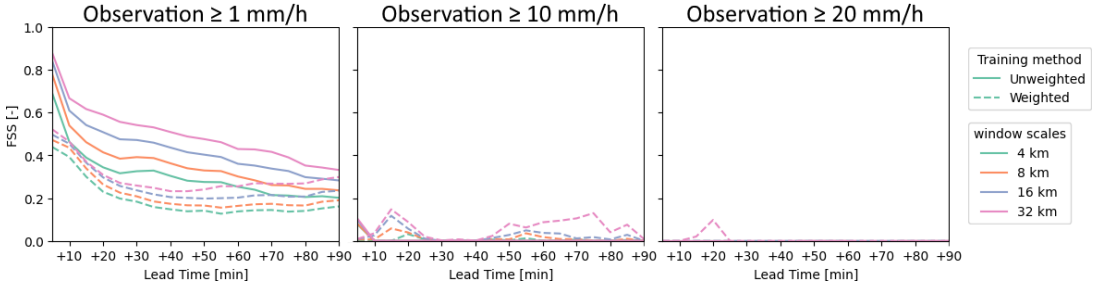

Figure B.2.: Average Fraction Skill Score (FSS) on 7 randomly sampled top 1% events from the test set for different threshold values on the DGMR model under the two different training strategies, higher is better.

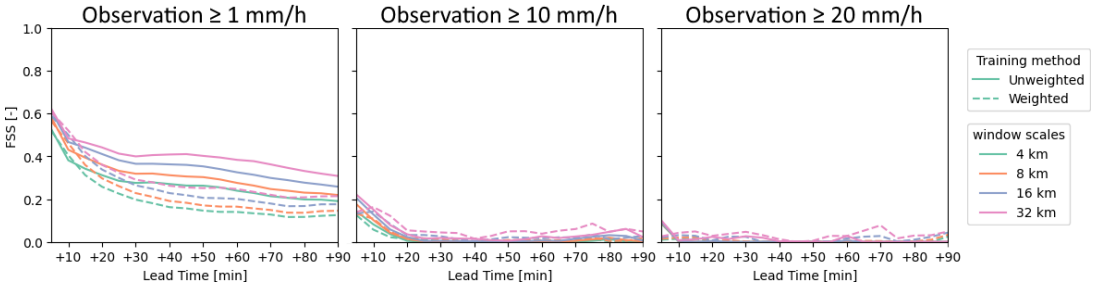

Figure B.3.: Average Fraction Skill Score (FSS) on 38 randomly sampled top 2% events from the test set for different threshold values on the DGMR model under the two different training strategies, higher is better.

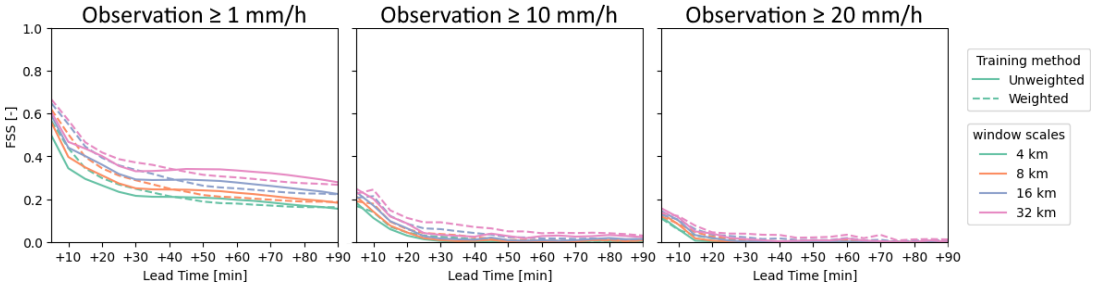

Figure B.4.: Average Fraction Skill Score (FSS) on 177 randomly sampled top 5% events from the test set for different threshold values on the DGMR model under the two different training strategies, higher is better.

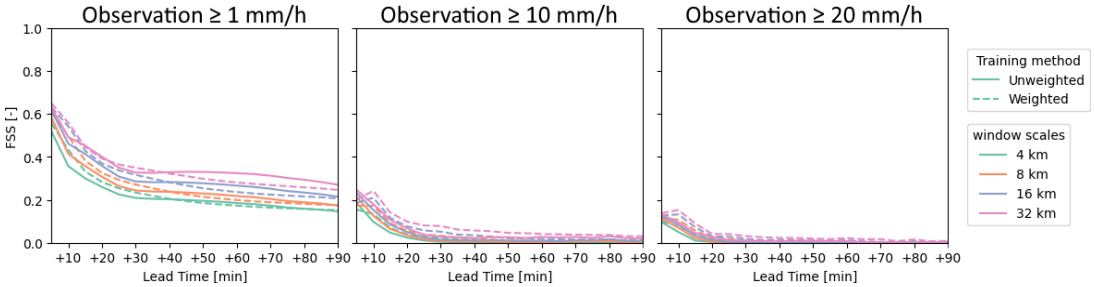

Figure B.5.: Average Fraction Skill Score (FSS) on 454 randomly sampled top 10% events from the test set for different threshold values on the DGMR model under the two different training strategies, higher is better.

C

Test events in linear scale

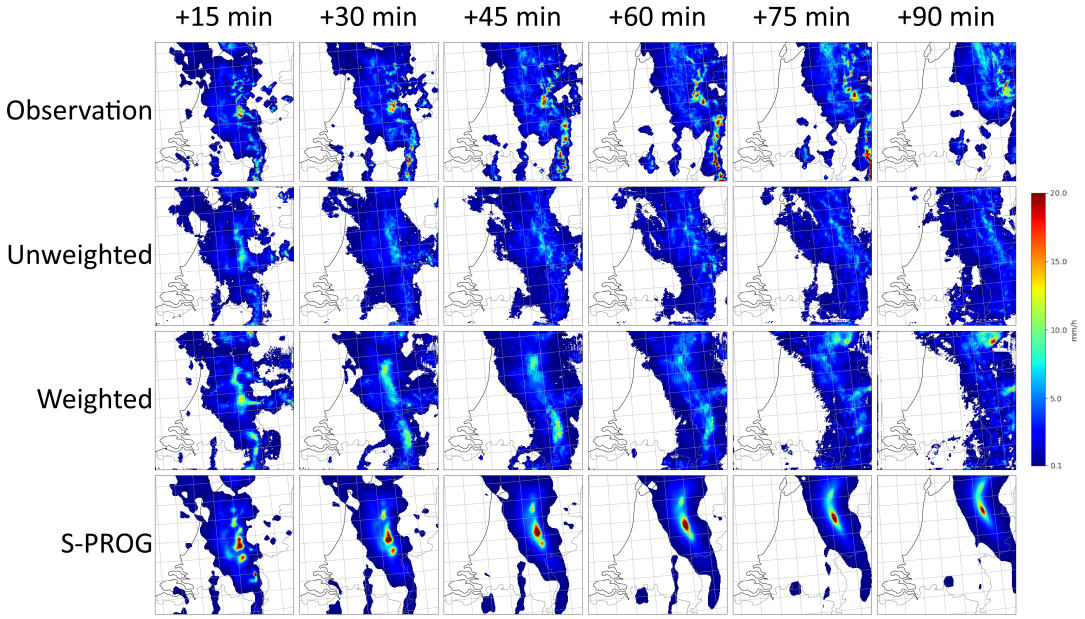

Figure C.1.: Observation and unweighted and weighted DGMR and S-PROG predictions for Event 1, at t_0 2022-05-19 12:05, plotted with a linear scale.

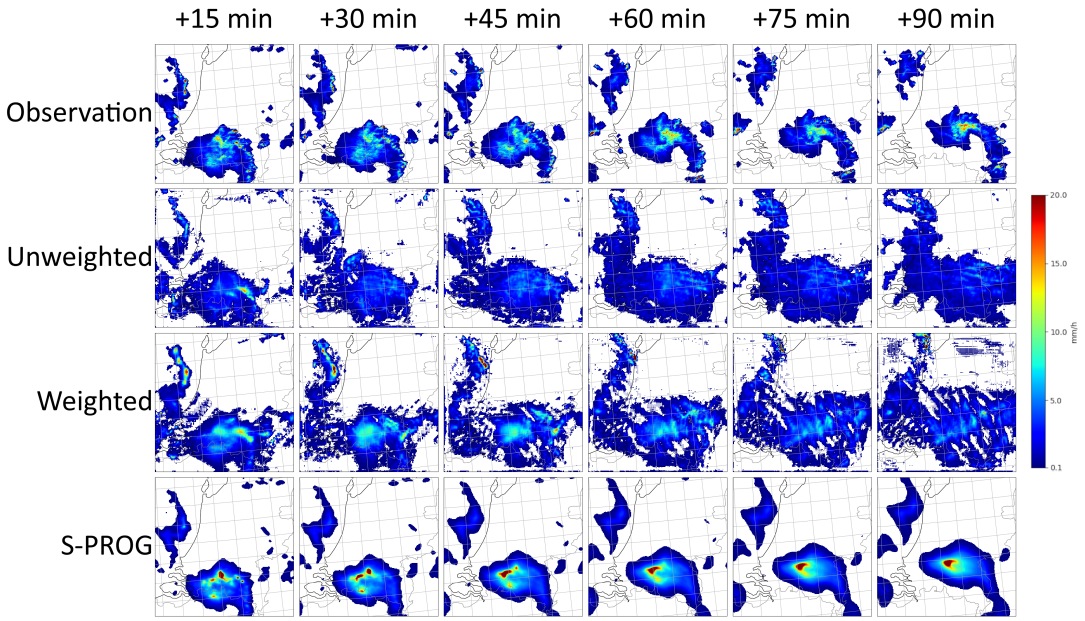

Figure C.2.: Observation and unweighted and weighted DGMR and S-PROG predictions for Event 2, at t_0 2022-09-08 20:25, plotted with a linear scale.

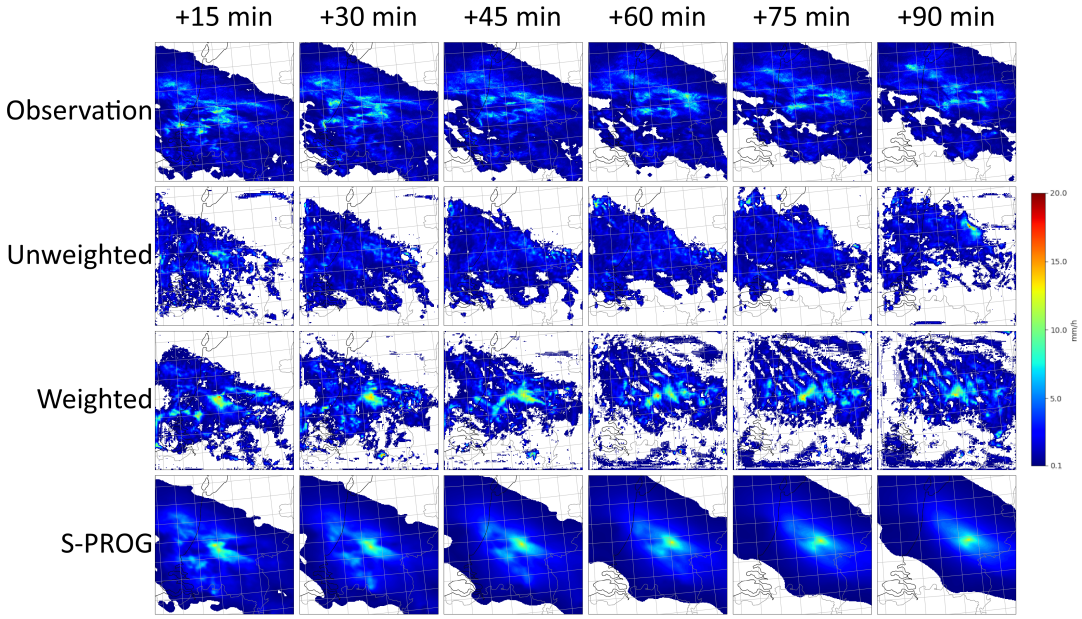

Figure C.3.: Observation and unweighted and weighted DGMR and S-PROG predictions for Event 3, at t_0 2022-12-23 12:20, plotted with a linear scale.

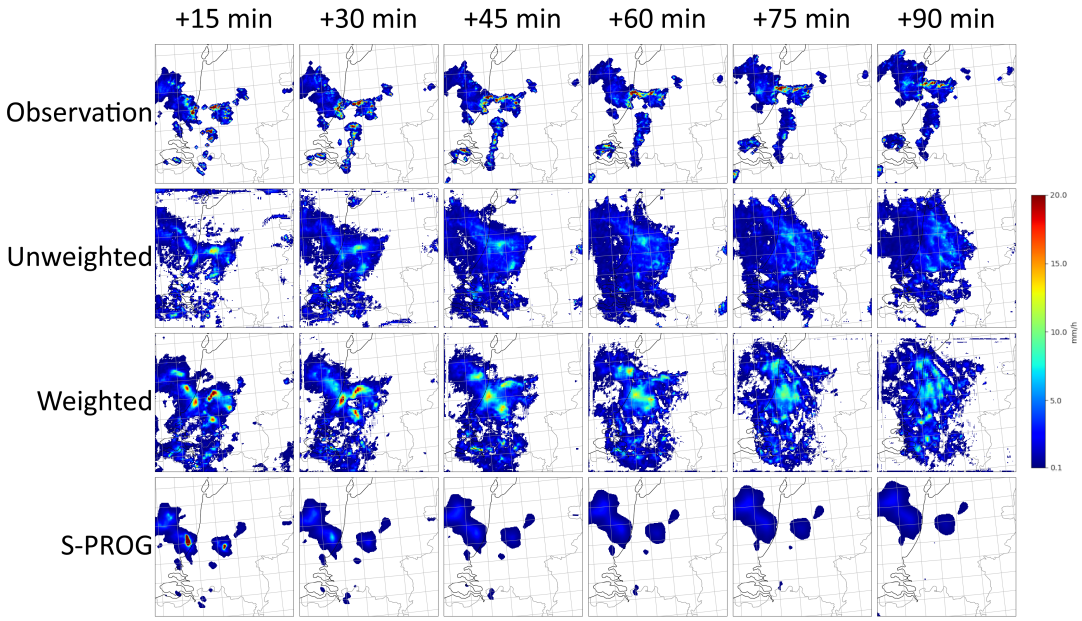

Figure C.4.: Observation and unweighted and weighted DGMR and S-PROG predictions for Event 4, at t_0 2022-08-17 01:55, plotted with a linear scale.

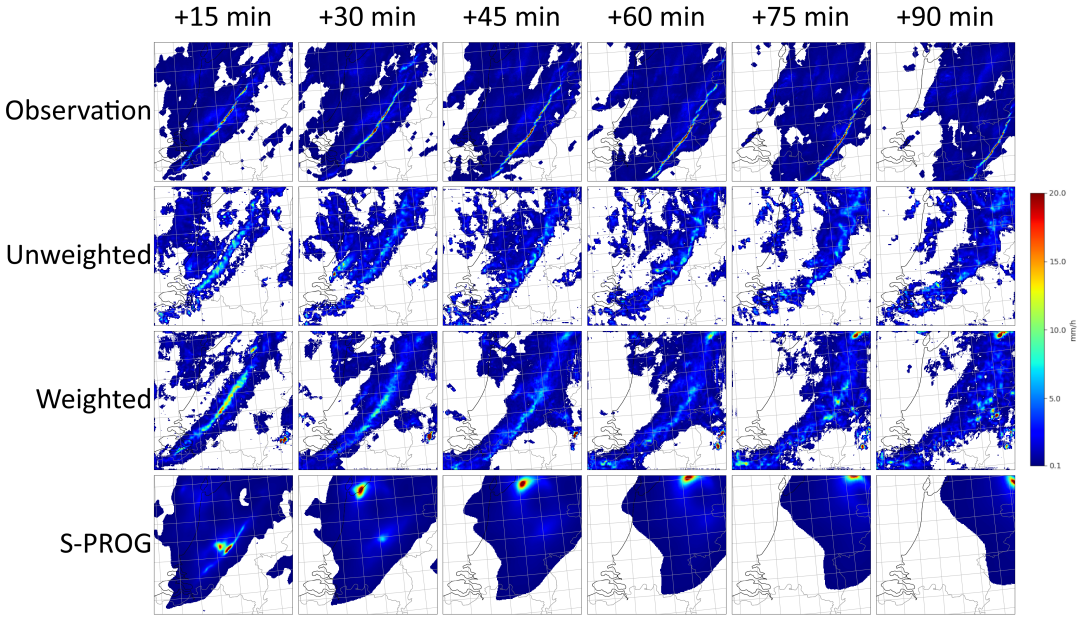

Figure C.5.: Observation and unweighted and weighted DGMR and S-PROG predictions for Event 5, at t_0 2022-02-20 20:55, plotted with a linear scale.

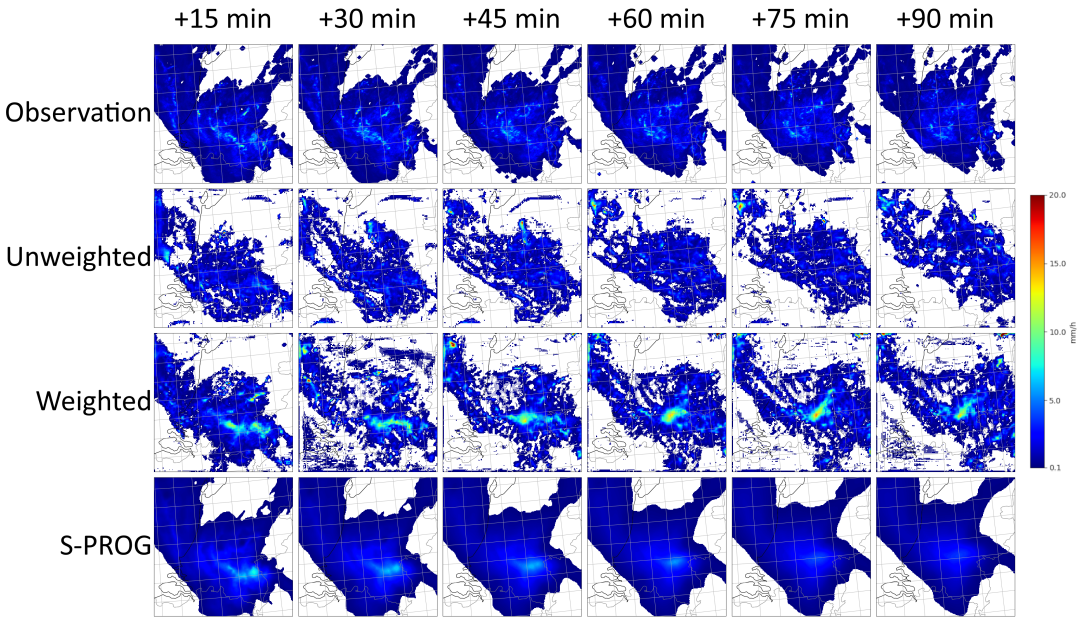

Figure C.6.: Observation and unweighted and weighted DGMR and S-PROG predictions for Event 6, at t_0 2022-06-05 15:00, plotted with a linear scale.

D

Original DGMR overview

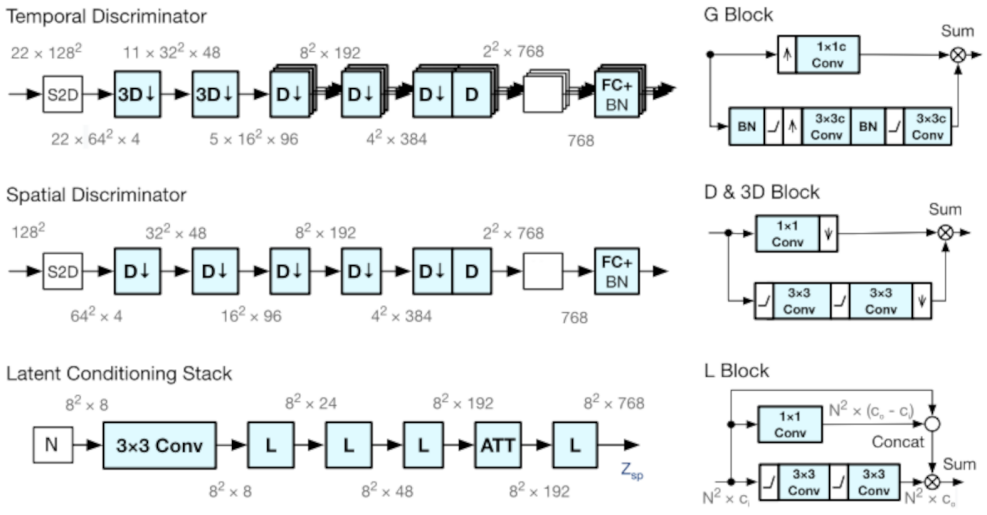

Figure D.1.: Original discriminators, Latent Stack and the architecture of the Generator block, Downsampling block and Latent block used in DGMR. Image taken from Ravuri et al. (2021).

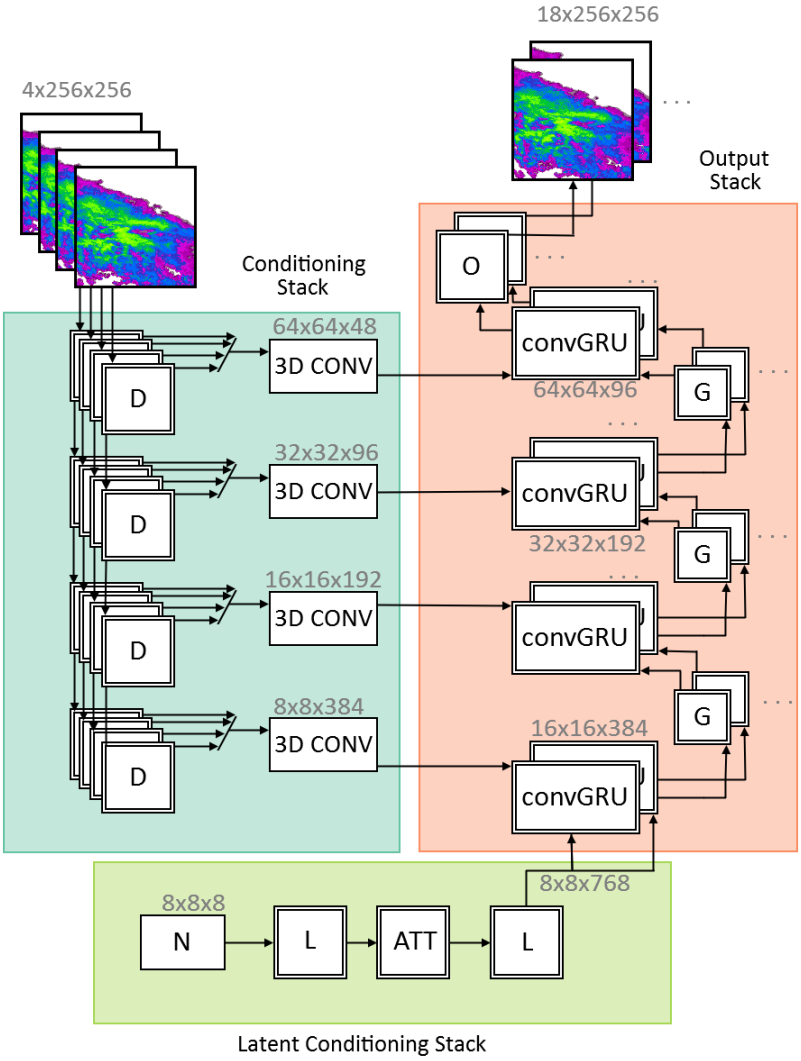

Figure D.2.: Schematic overview of the original DGMR Generator with the Conditioning Stack where the 4 input frames of size 256×256 are processed, the Latent Conditioning Stack where noise from a Gaussian distribution is processed and fed towards the Output Stack where 18 output frames of size 256×256 are generated. Image taken from Elsmann (2023) with further details of the blocks in Figure D.1.

E

Echo Top Height

E.1. Echo Top Height data

The KNMI radar systems make 14 scans with elevation angles of 0.3, 0.4, 0.8, 1.1, 2.0, 3.0, 4.5, 6.0, 8.0, 10.0, 12.0, 15.0, 20.0 and 25.0 degrees. The Echo Top Height (ETH) is then estimated by taking the maximum height above the earth's surface in kilometers with a reflectivity exceeding 7 dBZ, corresponding to 0.1 mm/h. The two radar images are then combined by taking the maximum ETH of the two images per pixel (Beekhuis and Holleman, 2008). The resolution of the product is 1×1 km and it can be downloaded at the KNMI data platform and starts on January 1, 2008, at 00:00 UTC.

The detection threshold of 7 dBZ might cause detection of spurious ETH values which may originate from planes or reflections from the tropopause. Due to the limited amount of elevation scans, there are gaps in the detection of ETH for the high elevation angles, causing ring-shaped gradients of the ETH (Aberson, 2011). The ETH in combination with the precipitation product can provide a simplistic form of the 3D structure of a precipitation field by providing the elevation of the precipitation field.

E.2. Precipitation rate - ETH analysis method

The ETH are only taken between 1 and 15 km, the lower boundary to prevent too much ground clutter and the top boundary as the cloud top height can reach up to 15 km during the summer, but higher values may be errors (Aberson, 2011). The precipitation rates are then plotted against the ETH for the test events from Section 4.4.4 in a pixel-wise manner over the entire duration of the event for the research domain as shown in Figure 4.5. The precipitation rate - ETH pairs are then binned on the precipitation rates, from 0.1 to 0.3 mm/h, 0.3 to 1.0 mm/h, 1.0 to 3.0 mm/h, 3.0 to 10.0 mm/h, 10.0 to 30.0 mm/h and 30.0 to 100.0 mm/h, as 100 mm/h is the highest occurring precipitation rate in these events. For each bin the distribution of the occurring ETH is plotted using violin plots.

E.3. Precipitation rate - ETH analysis results

In the ETH images from Figures E.1 and E.3 the ring-shaped gradients can be seen that are caused by the gaps in the detection for high elevation angles. These could pose an issue when the ETH is used as an input for a nowcasting model in situations where high ETH's occur. From all scatter plots showing the precipitation rate against the ETH as well as the violin plots (Figures E.2, E.4, E.6, E.8, E.10 and E.12 it can be seen that the measured ETH in general increases with precipitation rate. Especially the minimum ETH occurring at each precipitation rate shows a clear increase.

Since this analysis is comparing the precipitation rate and ETH on the same time step, it can be said that higher ETH's in the event coincide with higher precipitation rate on average on the same time step. Further research is required to see if ETH can give information on precipitation rates at further lead times.

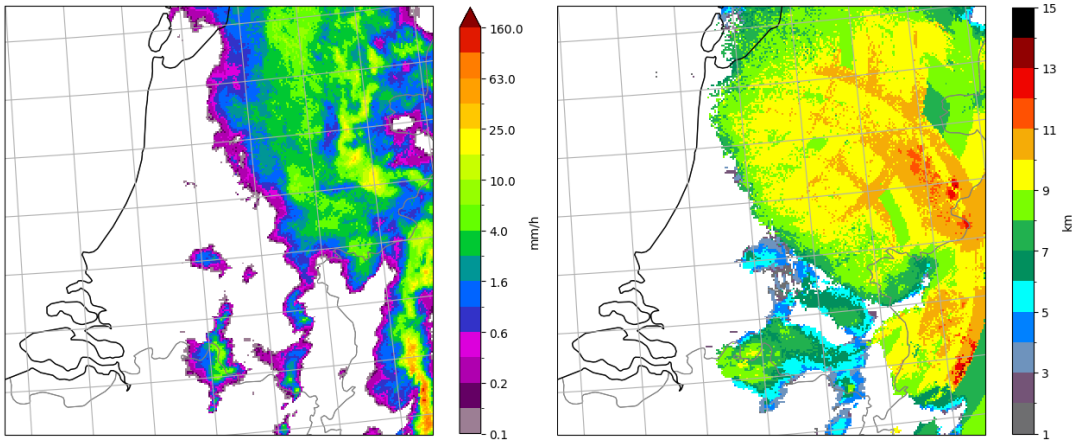

Figure E.1.: The Precipitation rate (left) and ETH (right) from test event 1 at 2022-05-19 13:15 UTC.

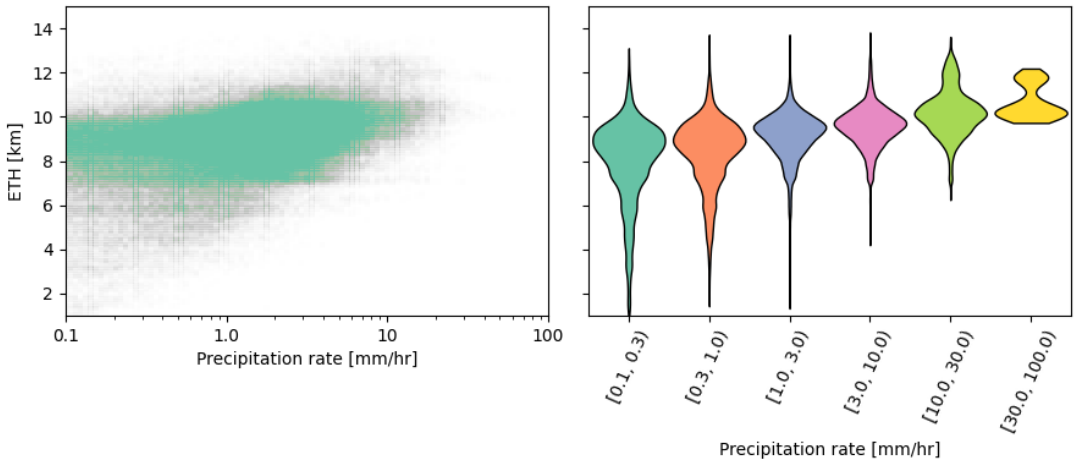

Figure E.2.: The precipitation rate plotted against the ETH for the entire event duration within the research domain (left) and the violin plots for the binned pairs (right) for test event 1 at t_0 2022-05-19 12:05 UTC.

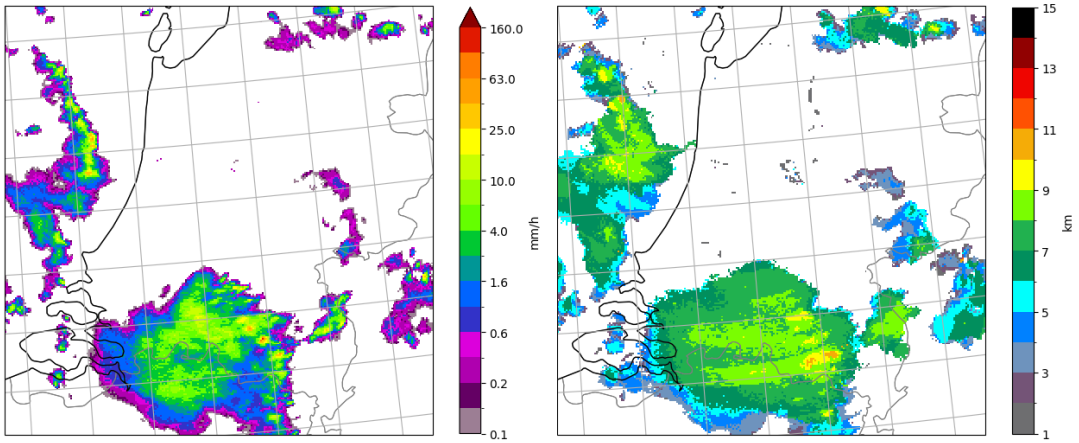

Figure E.3.: The Precipitation rate (left) and ETH (right) from test event 2 at 2022-09-08 20:25 UTC.

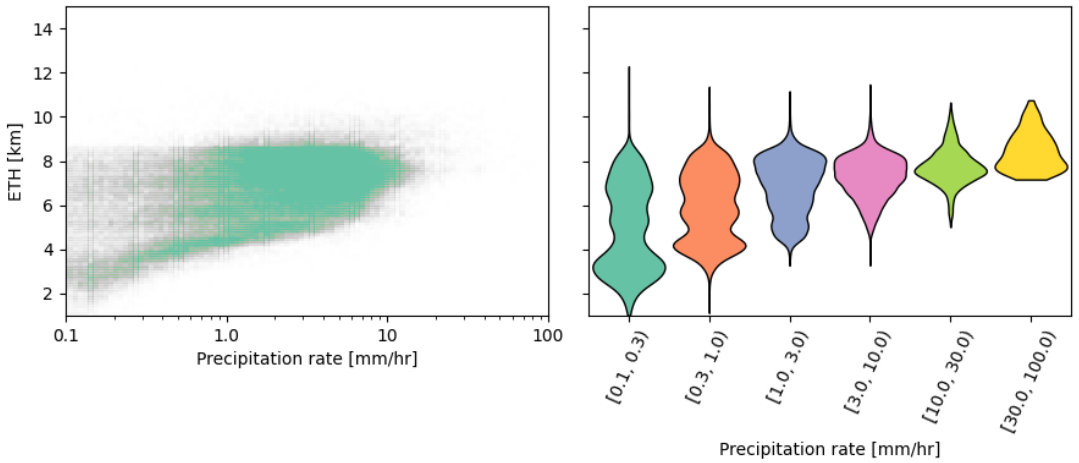

Figure E.4.: The precipitation rate plotted against the ETH for the entire event duration within the research domain (left) and the violin plots for the binned pairs (right) for test event 2 at t_0 2022-09-08 20:25 UTC.

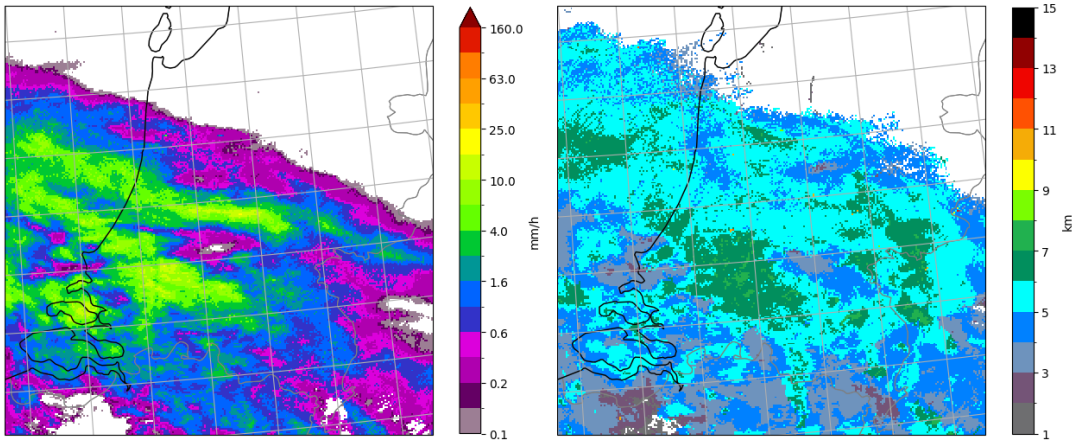

Figure E.5.: The Precipitation rate (left) and ETH (right) from test event 3 at 2022-12-23 12:05 UTC.

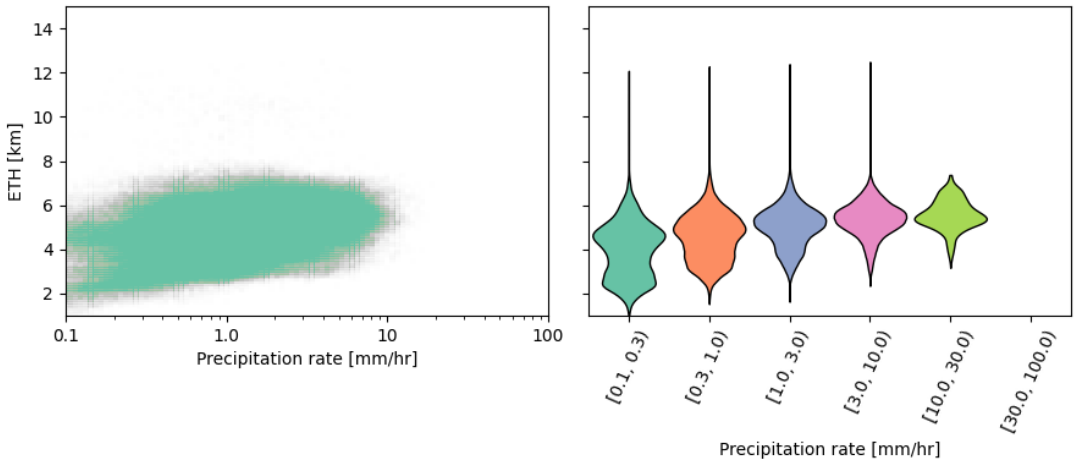

Figure E.6.: The precipitation rate plotted against the ETH for the entire event duration within the research domain (left) and the violin plots for the binned pairs (right) for test event 3 at t_0 2022-12-23 12:20 UTC.

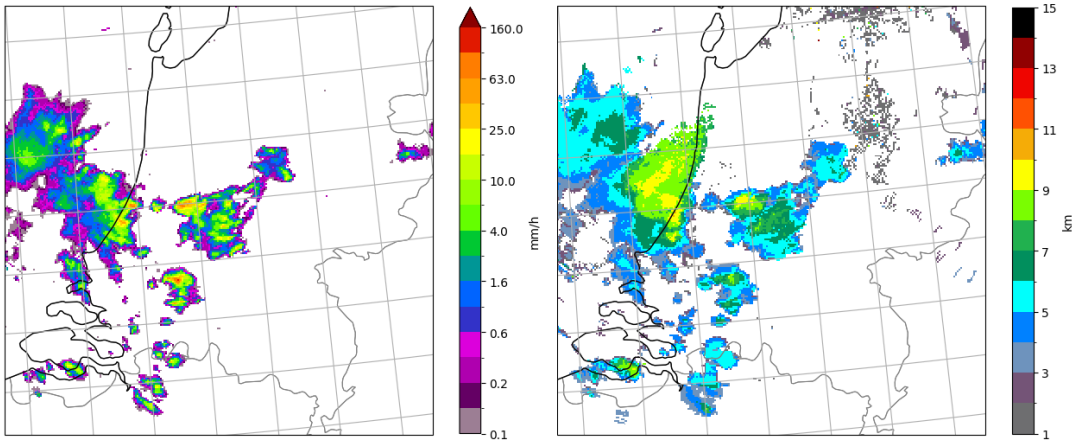

Figure E.7.: The Precipitation rate (left) and ETH (right) from test event 4 at 2022-08-17 02:05 UTC.

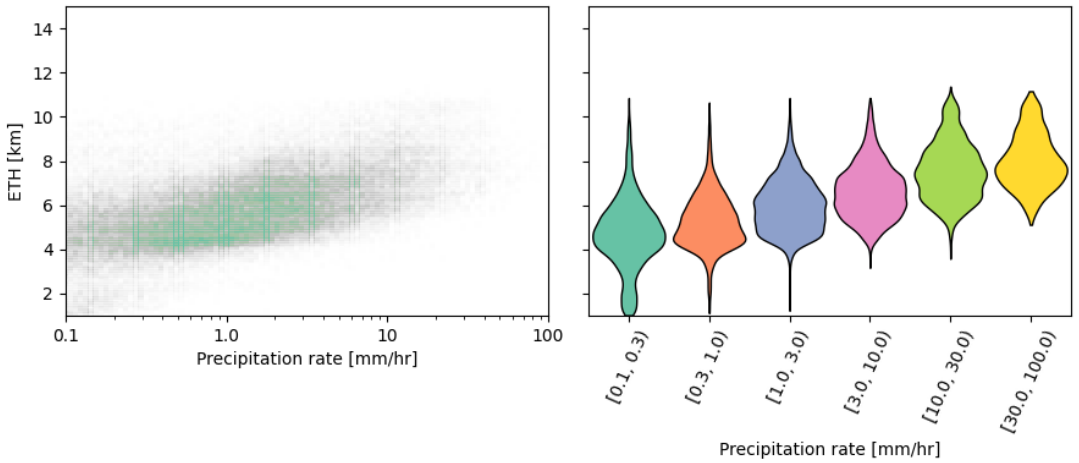

Figure E.8.: The precipitation rate plotted against the ETH for the entire event duration within the research domain (left) and the violin plots for the binned pairs (right) for test event 4 at t_0 2022-08-17 01:55 UTC.

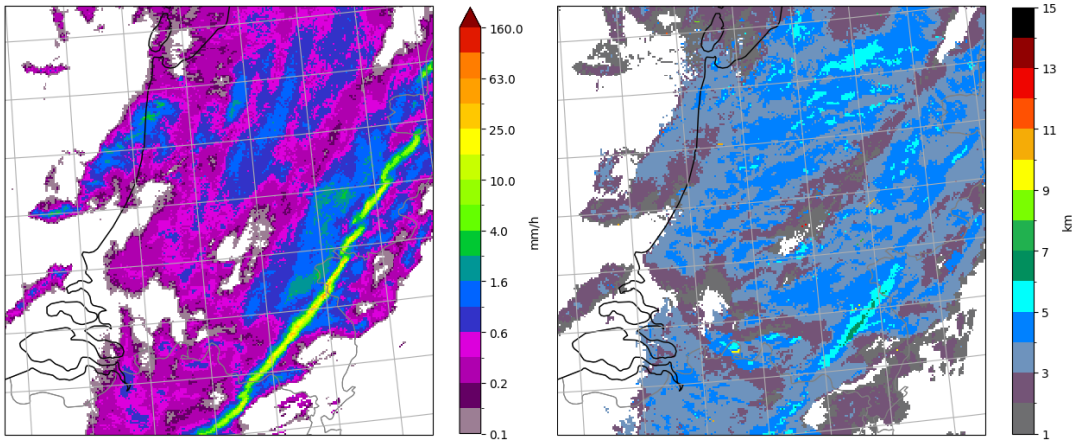

Figure E.9.: The Precipitation rate (left) and ETH (right) from test event 5 at 2022-02-20 22:00 UTC.

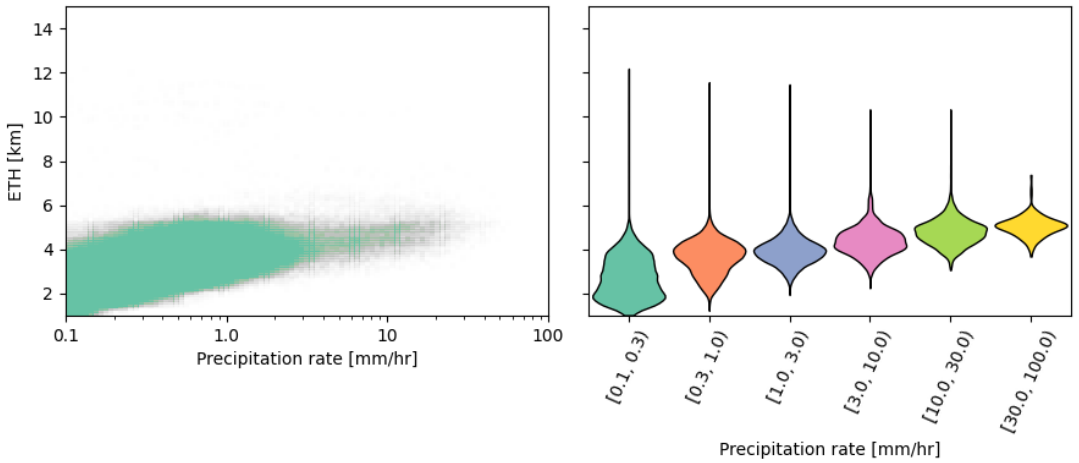

Figure E.10.: The precipitation rate plotted against the ETH for the entire event duration within the research domain (left) and the violin plots for the binned pairs (right) for test event 5 at t_0 2022-02-20 20:55 UTC.

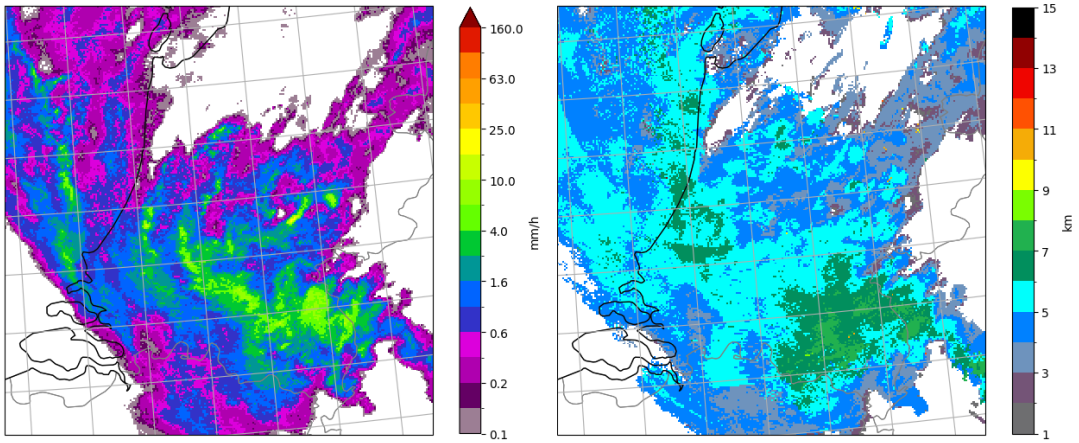

Figure E.11.: The Precipitation rate (left) and ETH (right) from test event 6 at 2022-06-05 15:05 UTC.

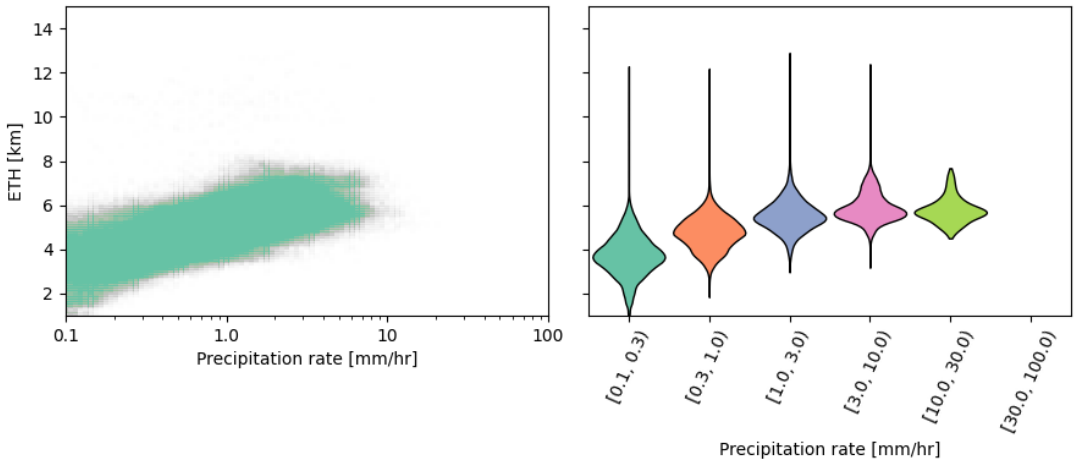

Figure E.12.: The precipitation rate plotted against the ETH for the entire event duration within the research domain (left) and the violin plots for the binned pairs (right) for test event 6 at t_0 2022-06-05 15:00 UTC.

Colophon

This document was typeset using LaTeX, using the KOMA-Script class `scrbook`. The main font is Palatino.