Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Masterarbeit am Institut für Meereskunde

Machine Learning-driven Infilling of Precipitation Recordings over Germany

Danai Filippou

Hamburg, January 2024

Supervisors:
Prof. Dr. Johanna Baehr
Dr. Christopher Kadow

Abstract

In this work I develop an AI/ML methodology that uses sparse meteorological observations to construct km-scale precipitation fields of weather radars over Germany. Radars provide data with high spatial (1 km^2) and temporal resolution (5min - 1hr), however the available radar dataset in Germany is limited to the 2000s onwards. The proposed methodology employs a well-established AI/ML model, known for its proficiency in infilling missing climate information. The model adopts the U-Net architecture with partial convolutional layers and employs a loss function which is specialized for infilling tasks. Various experiments are conducted, integrating different meteorological variables — precipitation, wind, pressure, and temperature from station measurements — along with temporal information through a multiple input timestep training setup. The models are evaluated against two reference datasets, namely the ground truth data RADKLIM and the gridded daily precipitation dataset HYRAS, using metrics that assess various aspects of model performance and accuracy. The AI/ML model shows remarkable accuracy, particularly when multiple input timesteps are incorporated. It demonstrates reasonable agreement with reference datasets in both spatial and temporal precipitation patterns, as indicated by the metrics. This best performing model ingesting multiple timesteps of station data, is then selected to reconstruct a 1km^2 precipitation field over Germany. I recreate hourly timesteps corresponding to intense precipitation events before and after the year 2000. Specifically, I examine the flood in the summer of 2021, intense rainfall events in May 2011, and the Pentecost flood event in 1999. The AI/ML model accurately reconstructs the precipitation that caused the 1999 flood event, demonstrating its physical plausibility and effectiveness in capturing past events. In conclusion, this innovative approach extends radar-based precipitation fields further back in time, providing insights into regional weather and climate that were previously inaccessible. The comparison between the highly resolved precipitation field and existing reanalysis products, which are computationally expensive, highlights the effectiveness of the trained AI model. The results are discussed in the context of related literature, including event reports by the German Weather Service, and contribute to our understanding of the regional past climate.

Acknowledgments

First and foremost, I would like to express my gratitude to Dr. Christopher Kadow for his guidance throughout this endeavor. Chris, your help has not only been significant to work aspects but also to many personal matters. Thank you for our countless constructive discussions, for always being available, patient and supportive to me and for all the opportunities you have given me. I would also like to thank Prof. Dr. Johanna Baehr for kindly agreeing to be my co-supervisor and for her consistent support throughout my studies in Hamburg, particularly during my master thesis work.

A special acknowledgment goes to the entire CLINT group at the DKRZ. Being part of this incredible group has been an amazing experience, and I will genuinely miss Johannes' beer and our Friday meetups. I am particularly grateful to Étienne Plésiat for our long and fun meetings. This work would not have been possible without his support.

My heartfelt thanks go to my parents, to whom I dedicate this work. I miss you every day, thank you for always supporting my decisions no matter in which part of the world they lead me and for always urging me to follow my dreams. I could not have done this without you.

Here in Hamburg, I would like to thank all my friends and colleagues from the university. Special thanks go to Nina; thank you for being an amazing friend, for the delicious food we cook, and for your support, no matter how far away we live from each other. Many thanks also to Katja for all the fun times we have spent together, looking forward to many more! Of course, thank you to all the friends from back home, especially to Despina, Vivi and Fotis, for our super long phone calls. I can't wait for the time we finally meet again.

Last but certainly not least, I express my deepest thanks to Görkem for his enduring love and unwavering support. Thank you for always being there for me and for believing in me more than anyone else.

Table of Contents

1. Introduction

Precipitation, in its different forms - from rain to hail or snow, plays a substantial role in shaping Earth's ecosystems and greatly affects human lives. It is a fundamental driver of temperature, humidity, and atmospheric conditions, and its study is essential for accurate weather forecasting. Intense precipitation events such as heavy rainfall or snowmelt can lead to natural hazards like floods and landslides. Such an event occurred quite recently, in the summer of 2021 in Western Germany: over the course of two days, from 14 to 15 July 2021, more than 180 lives were claimed and over 40,000 people were affected by a series of devastating floods (Junghänel et al. 2021). This catastrophe was the result of extensive rainfall during the whole summer of that year, as well as the influence of a slow-moving large summer storm system named 'Bernd' which was amplified in size and moisture from climate change. Precipitation levels reached up to 150 mm within 48 hours before the floods occurred (Fig. 1.1). As documented in the official reports by the German Weather Service (Deutscher Wetterdienst, DWD) the excessive amount of water caused ground saturation and overflowing of even the smallest of rivers in the German countryside, that led to the floods.

The 2021 summer flood was a well-documented case. The availability of high-quality radar data played a crucial role in accurately capturing and understanding this catastrophic event. However, as we delve deeper into the past, access to comprehensive and high-resolution datasets becomes more and more limited, while extreme events were of course always present. In May 1999, saturation of soil following extensive rainfall and snowmelt led to the 1999 Pentecost flood (Fuchs et al. (1999)). In 1997 after a series of heavy rain events in eastern Europe, the river Oder overflowed, resulting in casualties and damages across three countries including Germany (Bissolli et al. 2021). In the year 1962

more than 300 lives were claimed in a devastating flood that hit Hamburg (Mauch et al. 2012). Accurate historical data from these periods are extremely important when studying changes in the frequency and magnitude of such events. Moreover, reliable data from the past give us the opportunity to better analyze the regional climatology and assess present and future climate trends in precipitation. The results of these statistical analyses can be used to improve disaster preparedness and infrastructure planning, ultimately contributing to informed decision-making and policies related to climate adaptation and mitigation.

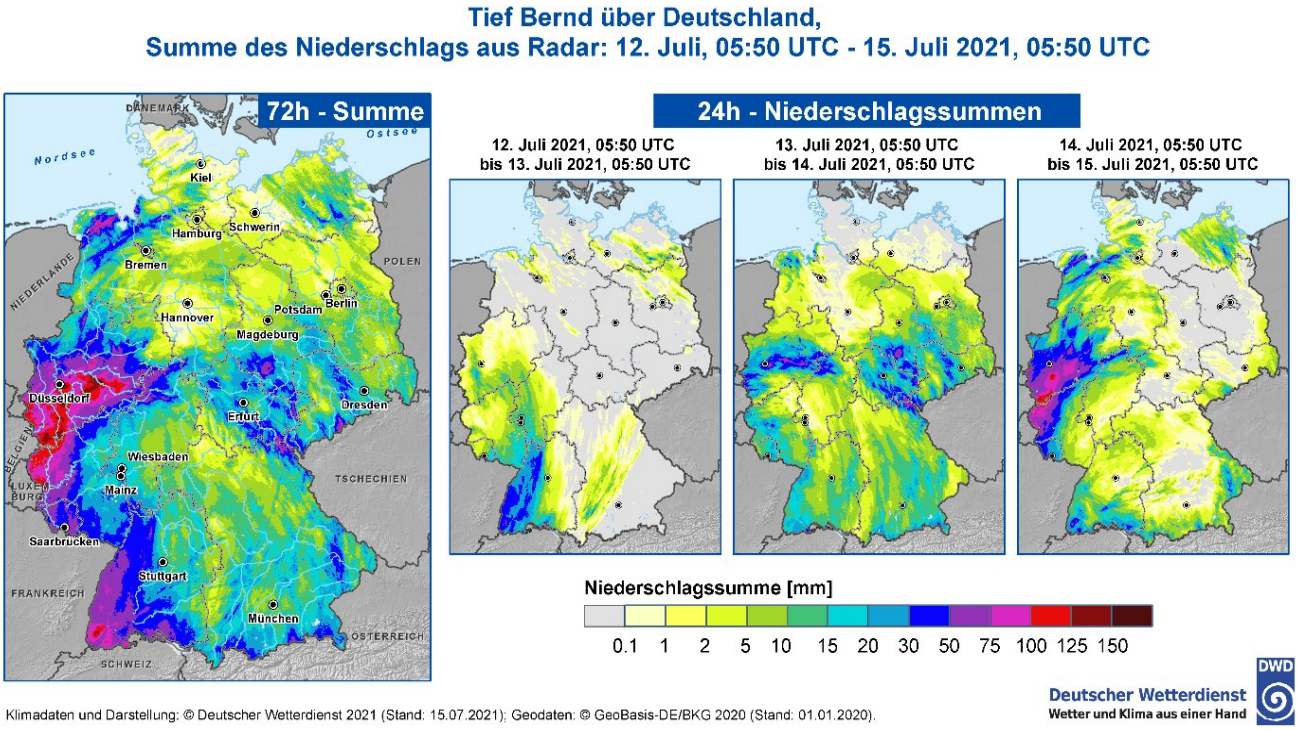


Figure 1.1: Precipitation levels analysis based on RADOLAN for the 24 h and 72 h until 15.07.2021 05:50 UTC (07:50 CEST). Source: DWD, Hydrometeorology (Junghänel et al. 2021)

Precipitation is a challenging field of study because it is highly variable in space and time. This means that its distribution and intensity rapidly change across various geographical locations and chronological periods. In terms of space, precipitation patterns can differ significantly from one place to another, even within relatively short distances. For instance, a neighboring town or region might experience vastly different amounts or types of precipitation within the same time frame. This spatial variability can be influenced by geographic features and proximity to water bodies. Regarding time, precipitation can vary not only from season to season but even within a single day or hour. Rainfall might be intense during one part of the day and then stop abruptly. This temporal variability is driven by various factors, including daily and seasonal weather cycles, climate

patterns, and atmospheric conditions (Cristiano et al. 2017).

Highly resolved data is crucial in order to capture the complex dynamics of precipitation. As further outlined in Chapter 2, the availability of such data can be limited for regional scales. Particularly in Germany, we acquire precipitation data from three sources: stations, reanalysis products and radars. Meteorological stations provide sporadic precipitation recordings, lacking continuous spatiotemporal coverage. However, they are the longest record of precipitation time series that we have at our disposal. On the other hand, reanalysis products can give us this spatial coverage, but they are computationally expensive to produce, and usually the datasets that provide longer time series (e.g. ERA5) have global coverage. As a result, they have a coarse resolution when we look at regional scales within country borders. Finally, radar-based precipitation monitoring grants an improved resolution in depicting precipitation patterns both spatially and temporally, making it a much better solution than stations and reanalysis. The main drawback however is that the radar network and its associated Quantitative Precipitation Estimates (QPEs) products have been operational in Germany only since 2001 (Müller et al. 2019). Looking back to the past before the 2000s, stations were one of the few precipitation data sources that existed with an hourly resolution. This leaves a gap in scientific knowledge: before the 2000s, we are lacking a highly resolved precipitation field with a fine temporal and spatial resolution, similar to what a radar can provide.

Various methods are traditionally used to fill such gaps in climate science for multiple purposes. For instance, data assimilation is a process used to integrate data derived from different sources (e.g., weather stations, radars, satellites) with numerical models to obtain an accurate and complete representation of atmospheric variables. Ensemble data assimilation (EDA) methods, such as the ensemble Kalman filter, use slight variations in initial conditions to estimate model variable dependencies and uncertainties, ultimately determining the current atmospheric state (Lahoz et al. 2014). However, measurements for a single timeframe are not sufficient to produce high quality results. To address this issue, data assimilation cycles are used. For example, in runs where the the global ICON model is used, three-hour cycles are implemented, as described by the German Weather Service. Thus, data assimilation is very computationally expensive, both due to this iterative nature and the fact that a great amount of data is processed in each cycle.

Another technique that is often used to generate synthetic climate time series are Weather Generators (WGs). Weather Generators are models that simulate marginal distribution and temporal dependence of meteorological variables, based on statistical

characteristics of observational data of a location of interest. One of the first and most widely used WG was proposed by Richardson et al. in 1981. In that work, precipitation is the first variable to be generated by a Markov chain-exponential model in a daily resolution. Three more variables (solar radiation, maximum and minimum temperature) are then simulated by a multivariate model, based on their correlations with each other and with the wet or dry status of each day. WGs following similar approaches are called the Richardson-type generators and they have been used since the 1980s to extend the length of climate datasets, infill missing data and provide meteorological information for locations where data is sparse (Yin et al. 2020; Yang et al. 2005, New et al. 2002). They are different from numerical climate models because they focus on smaller spatial scales and only on certain meteorological variables at a time. In contrast, climate models replicate the conditions of the whole atmosphere, as well as its interactions with the rest of the components of the Earth system, like the ocean and land (Ailliot et al. 2015). In that sense, WGs are computationally cheap tools and they perform reasonably well in terms of reproducing averages of variables but as every method, they too have a series of drawbacks. Their statistical approach can oversimplify the underlying dynamics, leading to unrealistic representations of meteorology. Additionally, WGs often struggle to accurately simulate certain types of extreme weather events due to their reliance on historical statistical patterns. Capturing extreme precipitation events with limited historical occurrences can thus be a challenging task (Yang et al. 2005).

Kriging is another widely used geostatistical interpolation method in climate science (Oliver et al. 1990). Kriging applies the assumption that nearby points contribute more to the estimation than distant ones, taking into account the clustering of points. Clusters of data points have less influence on the predictions as they contain less unique information. This approach helps counter bias in the predictive results. Additionally, kriging provides estimates of uncertainty for each interpolated value. Finally, what makes it different than simpler methods, such as Inverse Distance Weighted Interpolation, Linear Regression, or Gaussian decay, is its reliance on the spatial correlation between sample points for interpolation. Rather than making assumptions about the spatial distribution, kriging leverages the actual spatial arrangement of observed data points. It is a very successful method when it comes to interpolating variables with strong spatial autocorrelation. (Auchincloss et al. 2007). This means that the value of the variable at one location is closely related or correlated with the values of the same variable at nearby locations. However, when it comes to precipitation, this kind of spatial autocorrelation is not typically observed. Previous works have shown that variations of kriging can be successful

when handling precipitation, when more variables are incorporated. In the work by Haberlandt (2005), kriging with external drift (KED) and indicator kriging with external drift (IKED) are used, along with a dense daily precipitation network combined with hourly rain gauge and radar measurements. This approach yields better results than other reference methods like Nearest Neighbor interpolation or Inverse Square Distance Weighting (IDW). However, it is worth noting that despite its effectiveness, kriging, especially when applied with multiple variables and high-resolution data, can be very computationally expensive (Haberlandt et al. 2005).

In recent years, Machine Learning (ML) has emerged as a promising alternative in the realm of climate science. ML techniques offer numerous advantages over traditional methods like kriging and data assimilation, especially when it comes to computational efficiency. ML models can leverage the parallel processing capabilities of Graphics Processing Units (GPUs), which enable the simultaneous execution of numerous calculations, drastically speeding up both training and inference processes. Additionally, many ML frameworks are optimized for numerical operations and incorporate specialized libraries and hardware, ensuring efficient computation and resource management. These characteristics help the ML models to quickly process vast amounts of data, making them well-suited for handling the complex and high-dimensional datasets often encountered in climate science.

ML methods have demonstrated their capability to capture intricate patterns and relationships within the data, making them a valuable tool for both prediction and interpolation tasks. In recent works they have even proven more accurate and fast than traditional Numerical Weather Prediction (NWP) models (Kurth et al. 2023) and they have exhibited remarkable skill in infilling missing climate information (Kadow et al. 2020). In their work, Kadow et al. (2020) infill temperature grids by using a machine learning model that was initially developed for image inpainting of irregularly shaped holes (Liu et al. 2018). Their approach manages to infill the El Nino event of 1877, even though very few data was available in that period of time. Moreover, when compared to methods like kriging and Principal Component Analysis (PCA), it yields more accurate results. Consequently, this particular machine learning method has the potential of beating traditional interpolation methods, when it comes to both speed and accuracy. In this Master Thesis, I apply this approach to radar and station measurements acquired from the German Weather Service (Deutsches Wetterdienst, DWD) with the final goal of reconstructing precipitation fields over Germany. I take advantage of this ground-breaking Machine Learning technique to combine the high spatial resolution a radar provides and

the long time series recordings of weather stations. The final result is a highly resolved precipitation field in both space and time for Germany in a period spanning before the 2000s.

In the second chapter of this Master Thesis I describe in detail the Machine Learning model, along with the radar and station data that were used, and the training workflow. Chapter 3 presents the evaluation of the trained models using different metrics. I then use the best performing model in Chapter 4 to reconstruct 1997 and 2012, constructing artificial precipitation fields that resemble radars in the past and present. I compare the model's output to the present-day data that are available. Chapter 4 also includes the result analysis and discussion, according to an event-based approach. Finally, the thesis concludes with Chapter 5, where a summary and final remarks are made, along with the future perspectives of this project.

2. Background and Methodology

The first section of the chapter outlines current approaches to infilling precipitation data. In the second and third section I give more details regarding the radar and station data used for this work. The fourth and fifth section introduce the model and the training setup. Finally, the sixth section serves as an overview of the reference datasets used throughout the thesis.

2.1 State-of-the-Art Precipitation Infilling Methods

There are numerous research works focusing on infilling missing precipitation data, with the majority of studies concentrating on monthly or daily frequency, in various regions around the world. Two kinds of techniques are conventionally used: statistical methods and machine learning. Until recently, geostatistical methods were more commonly used. For example, in the work by Bárdossy et al. (2014), Gaussian and unsymmetrical v-copula methods are compared against traditional kriging and External Drift Kriging (KED). The study focuses on daily precipitation measurements of 41 years in three areas of Germany and concludes that the copula methods outperform the other methods, both in terms of bias and uncertainty. Verworn et al. (2011) use KED with daily station and radar data along with topography information in order to interpolate hourly data in the region of Ummendorf in northern Germany. Their study shows that for winter stratiform events, data from daily weather stations are sufficient, while the use of radars enhance the interpolation accuracy when it comes to convective summer events. The authors however point out that their findings are greatly dependent on the topography of the considered regions and a different approach is needed in locations with other topographies. Berndt et

al. (2018) investigate how the performance of various interpolation methods is affected by station density, temporal resolution, and spatial variation of the meteorological variables that are interpolated. Their study is focused on a region within 128km of the radar station in Hanover. The first general conclusion they reach is that interpolation performance is more closely correlated with temporal and spatial resolution of each variable and less dependent on the stations density. When comparing different interpolation methods, they find that for all climate variables ordinary kriging gives better results that the simple Nearest Neighbor and Inverse Distance Weighting approaches, except for precipitation, where the accuracies are comparable when a dense station network is in place. Moreover, the results for precipitation indicate that this was the most challenging quantity to interpolate due to its highly variable nature. Once again radars were a valuable asset and helped to increase interpolation accuracy for hourly resolutions.

In more recent works, Machine Learning (ML) became an alternative for precipitation infilling tasks. In the work by Londhe et al. (2015) Artificial Neural Networks (ANNs) are used to forecast daily precipitation values, given the rainfall amounts measured at 11 rain gauges scattered across the Pune District of India. A work by Coulibaly et al. (2007) focused on Canada, where 6 different types of Neural Networks are used to fill gaps in daily precipitation time series of 15 weather stations in the Gatineau watershed. A notable finding of that study is that two dynamical networks (the Recurrent Neural Network and the Time Delay Neural Network) produced less accurate results than the rest of the models, even though they incorporated temporal information. Militino et al. (2022) propose the use of Neural Networks, Random Forests and k-Nearest Neighbour methods against kriging to interpolate data of rain gauge stations in Spain, using only the geographical coordinates and the altitude of each station. They conclude that Machine Learning methods can be a convenient alternative to kriging since they provide predictions of comparable accuracy, while being less computationally expensive. In Europe, Moraux et al. (2019) conduct a study focused in Germany, Belgium and the Netherlands, where they use a Deep Learning (DL) model to forecast precipitation rates. They use rain gauge measurements from the three countries, as well as satellite data and an advanced encoder-decoder Convolutional Neural Network (CNN) that is originally used for semantic segmentation. By calculating a series of scores for their models' performances they conclude that their Deep Learning approach can combine the advantages of existing rain gauge interpolation and satellite data and yields considerably more accurate results than previous methods. Rojas-Campos et al. (2022) use Numerical Weather Prediction (NWP) models and various Deep Learning models to generate high-resolution precipitation maps from low-resolution NWP model output. The research highlights a marked improvement in the predictive accuracy of DL models when they integrate data from NWP models. Additionally, it emphasizes that optimizing the complexity level of the model's architecture is a key factor in enhancing its performance. Some studies compare the interpolation skill of statistical and ML approaches. Teegavarapu et al. (2017) use Inverse Distance

Weighting Method (IDWM) variations, ANNs and several benchmark methods to infill missing daily precipitation data at 53 rain gauges of South Florida, USA. They conclude that the Linear Weight Optimization method (LWOM), which is similar to a multiple linear regression, along with the ANN yield the best prediction scores.

While there has been substantial research focusing on infilling missing precipitation data, especially with the rise of Machine Learning as a promising approach, few of the existing studies have gone further than infilling daily temporal resolution. To the best of my knowledge, there have not been any recent study in Germany focusing on infilling past hourly precipitation data through Machine Learning techniques. My master's thesis aims to address this significant research gap.

2.2 Image Inpainting Methods

Image inpainting in computer science refers to the process of filling in missing or corrupted parts of an image. It is a fundamental task in image processing and computer vision and has numerous applications, including photo restoration, object removal and image editing. There are various techniques used for image inpainting, from traditional methods to more advanced deep learning approaches. Traditional methods use algorithms that consider the colors and textures of neighboring pixels, as well as structures within the image, to generate a complete output. However, these methods are computationally expensive and often inaccurate. For example, PatchMatch is an algorithm that primarily relies on finding the best matching patches from the surrounding area to fill in missing regions in an image (Barnes et al. 2009). However, it lacks the capability to understand the higher-level meaning and context of the image, relying more on statistical similarities. This can result in inpainted areas that fit well statistically but are contextually unrealistic. In recent years, image inpainting has benefited from the development of Convolutional Neural Networks (CNNs) (Cai et al. 2015) and Generative Adversarial Networks (GANs) (Yu et al. 2018). These deep learning models can learn complex patterns and textures from large datasets, enabling them to produce realistic images.

Liu et al. (2018) apply a robust Machine Learning model, able to infill large and irregular holes of missing data in images. Such an achievement is possible by using partial convolutions in every layer of their model, which is using the U-Net architecture. The model is trained,validated and tested on images from three well-established databases: ImageNet (Russakovsky et al. 2015), Places2 (Zhou et al. 2017) and CelebA-HQ (Karras et al. 2017). Initially, a complete image is masked using a binary mask that is applied through element-wise multiplication (Eq. 2.1, where X is the original image, M is the mask and X' is the masked image), in order to simulate a corrupted image. The mask indicates which pixels are valid and which pixels should be infilled. As it is a binary mask, it only contains two values: 0s for missing pixels and 1s for valid pixels. During the partial convolution process, the convolution operation is applied only to the valid pixels, as

indicated by Eq. 2.2, where W is the convolution filter weights and b is the corresponding bias (the term sum(1)/sum(M) serves as a scaling factor). This operation ensures that the network will only use valid values to predict the missing ones. In this way, the masked areas do not contribute to the convolution operation, thus preventing the integration of any incorrect information from the damaged regions. Another important aspect of the method is an automatic mask update mechanism in each layer. If a convolution operation covers at least one valid pixel, the center pixel in the mask (corresponding to the convolution's output location) is marked as valid for the next layer. This dynamic update is expressed through Eq. 2.3 and it allows the network to reconstruct the missing values. Given sufficient mask updates, the output of the network will be a complete image.

$$X' = X \circ M \tag{2.1}$$

$$x' = \begin{cases} W^T(X \circ M)\frac{sum(1)}{sum(M)} + b, & if\ sum(M) > 0 \\ 0, & otherwise \end{cases} \tag{2.2}$$

$$m' = \begin{cases} 1, & if\ sum(M) > 0 \\ 0, & otherwise \end{cases} \tag{2.3}$$

This method has already found several successful applications in the domain of climate science. Kadow et al. (2020) use the same model by Liu et al. (2018) to reconstruct global temperature data. They use reanalysis data from the 20th Century Reanalysis (20CR) dataset (Compo et al. 2011), as well as climate model output data from the Coupled Model Intercomparison Project Phase 5 (CMIP5) (Taylor et al. 2012). The Machine Learning model is trained on multiple ensembles of each dataset, and is validated on one ensemble that was excluded from the training dataset. In this application, the masks of missing values are extracted from the HadCRUT4 dataset which contains observational temperature data and a great amount of missing values (Morice et al. 2012). After successful evaluation on the test datasets, the trained models have been used to reconstruct the HadCRUT4 dataset. A comparison with other statistical methods such as kriging revealed that the deep learning technique was able to capture more intricate spatial patterns. In particular, it allowed the reconstruction of an El Niño event from July 1877 that was documented but absent form the original HadCRUT4 dataset.

In a related study, Meuer et al. (2022) extend the method by Kadow et al. to precipitation data. They further integrate a Long-Short Term Memory (LSTM) module and in doing so, they manage to better capture the temporal variability of precipitation. Additionally, they include an attention-based module that incorporates information from reanalysis data such as wind and temperature. In their research, they apply successfully their methods to the infilling of missing values from the RADOLAN dataset. Furthermore,

the study suggests that a combination of these advanced modules yields improved results compared to the original baseline method.

2.3 Radar Data in Germany

The data used in this study is derived from two main sources: radars and stations. Radars (short for Radio Detection and Ranging) operate by emitting electromagnetic signals in the atmosphere. When these pulses encounter particles of precipitation such as rain, snow, sleet, or hail, a fraction of the energy is scattered in all directions, with a small portion reflecting back toward the radar. Assuming that these hydrometeors are spherical small, the power of the reflected pulse P is related to the radar reflectivity factor Z according to the following equation:

$$P = Z\frac{C|K|^2}{l^2 r^2} \tag{2.4}$$

where C is a constant related to the radar characteristics (antenna gain, pulse width, wavelength, beam width and pulse duration), $|K|^2$ is the dielectric factor of the particles, r is the distance from the radar, and l^2 is the attenuation term cause by the two-way propagation in gases, clouds and precipitation (Sauvageot et al. 1994). By analyzing the time it takes for the echo to return and the strength of the received signal, the radar can determine the location, intensity, and movement of different types of precipitation. Figure 2.1 shows two examples of weather radar systems (Winterrath et al. 2017).

Radars started being used for weather applications in Germany roughly after the 1960s. Over time, the quality and capability of radar technology has significantly evolved. The DWD produces extensive reports that outline in detail the technological upgrades of the radar network (Winterrath et al. 2017). As of 2001, an important milestone in enhancing data quality has been the transition to Doppler radars, which exploit the Doppler effect to differentiate moving hydrometeors from stationary objects that could negatively impact precipitation measurements. A subsequent upgrade involved shifting from single to dual-polarization radars, which measure in two polarization planes, providing a more accurate assessment of the shape, phase, and size distribution of hydrometeors. A timeline of upgrades of the network are illustrated more in detail in Figure 2.2. Furthermore, some radar sites were relocated, especially those in urban areas, to reduce the disturbances caused by buildings (Winterrath et al. 2017). All of the 17 radar devices that currently comprise the main DWD network, as well as a research radar located in the Hohenpeissenberg Meteorological Observatory that is used for quality assurance, employ the dual-polarization Doppler technology. The full list of these radars along with a series of their main characteristics is provided by the DWD and is presented here in Table 2.1., while their exact location can be seen on the map on Figure 2.3. These

radars scan the atmosphere every 5 minutes, within a radius of 150km. Other radars with lower capabilities are operational as well, but are not as widely used.

(a) (b)

Figure 2.1: a) A covered weather radar located in Offenthal, Germany. The system is mounted on a tower that is 45m high and is protected by a radome. Four lightning rodes around it protect the technicians and the radar electronics (control unit and transmitter), that can be found in the interior of the tower. b) An unprotected radar system (right). The antenna can be moved in different angles. The receiver in this case is mounted behind the reflection system. The movement of the weather radar antenna ensures a high-quality scan of the atmosphere that provides high spatiotemporal resolution for all DWD radars. (Winterrath et al. 2017)

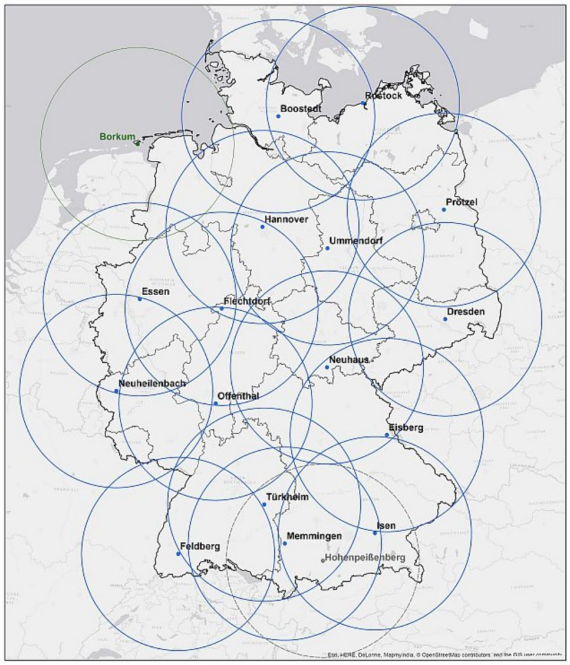


Figure 2.2: The complete radar network of the DWD, consisting of 16 ground radars for regular scanning of the atmosphere (blue), 1 Airport Surveillance Radar (ASR) in Borkum (green) and 1 radar for quality assurance purposes in the Hohenpeissenberg Meteorological Observatory (dashed gray). (Source: Radar Network - DWD)

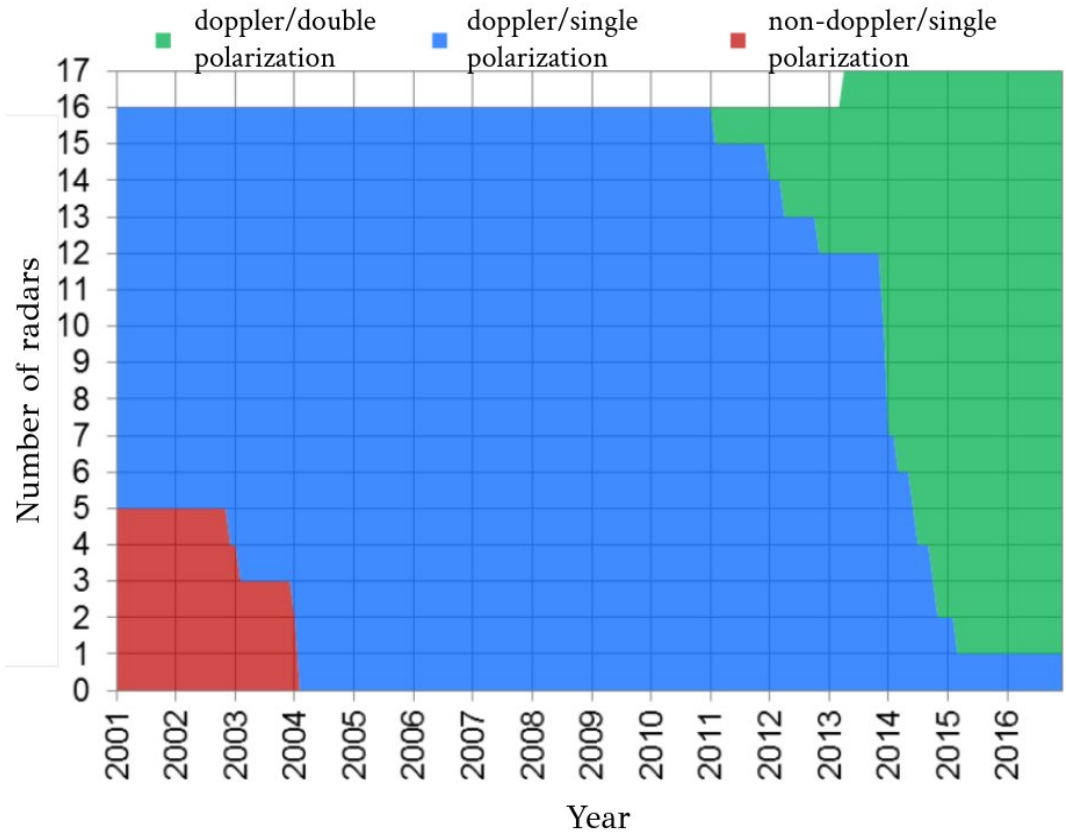


Figure 2.3: Diagram of the technical upgrades that the DWD radar network underwent in the period 2001-2016. Modified from Winterrath et al. (2017).

Location	WMO number	Geographical coordinates	Double Polarization Doppler
ASR Borkum (ASB)	10103	53° 33‘ 50,44‘‘N 6° 44’ 53,85‘‘E	unknown
Boostedt (BOO)	10132	54° 00' 15,8''N 10° 02' 48,8''E	23.01.2014
Dresden (DRS)	10488	51° 07' 28,7"N 13° 46' 07,1"E	17.03.2015
Eisberg (EIS)	10780	49° 32' 26,4''N 12° 24' 10,0''E	08.10.2014
Essen (ESS)	10410	51° 24' 20,3"N 06° 58' 01,6"E	11.04.2012
Feldberg (FBG)	10908	47° 52' 25"N 08° 00' 13"E	20.11.2012
Flechthof (FLD)	10440	51° 18' 40,3''N 08° 48' 07,2''E	12.11.2014
Hannover (HNR)	10339	52° 27' 36,3"N 09° 41' 40,3"E	29.07.2014
Isen (ISN)	10873	48° 10' 28,9"N 12° 06' 06,4"E	22.01.2014
Memmingen (MEM)	10950	48° 02' 31,7"N 10° 13' 09,2"E	03.04.2013
Neuhaus (NEU)	10557	50° 30' 00,4''N 11° 08' 06,1''E	10.01.2012
Neuheulenbach (NHB)	10605	50° 06' 34,8"N 06° 32' 54,0"E	27.03.2014
Offenthal (OFT)	10629	49° 59' 05,1''N 08° 42' 46,6''E	15.02.2011
Prötzel (PRO)	10392	52° 38' 55,2"N 13° 51' 29,6"E	23.01.2014
Rostock (ROS)	10169	54° 10’ 32,4‘‘N 12° 03’ 29,1‘‘E	11.06.2014
Türkheim (TUR)	10832	48° 35' 07,4"N 09° 46' 57,6"E	09.12.2013
Ummendorf (UMD)	10356	52° 09' 36,3"N 11° 10' 33,9"E	17.12.2013

Table 2.1: Description of the main radar network of the DWD (Source: Radar Network - DWD)

Even though radars provide high quality measurements with detailed resolution in space and time, they are an indirect measuring method (since rain rates and distributions are inferred from measured reflectivities) and they are prone to errors. Villarini and Krajewski (2009) provide an extensive overview of the sources of these uncertainties. One major factor of uncertainty is the attenuation of the radar beam, which occurs when the beam passes through heavy precipitation or encounters a wet radome. This attenuation leads to a decrease in reflectivity with increasing distance from the radar, resulting in an underestimation of precipitation intensity. The wet radome, in particular, can cause significant transmission loss, especially during moderate rainfall, and its impact is exacerbated by factors such as radome cleanliness. Another critical source of error is radar miscalibration, which refers to inaccuracies due to changes in the radar constant C. This constant can be affected by the deterioration of various radar components and thermal effects, leading to consistent errors across the radar's coverage area until recalibration. Ground clutter, caused by radar signal scattering from nearby objects like buildings and terrain, creates echoes that are not related to meteorological phenomena, introducing noise into the data. Beam blockage is particularly problematic in mountainous areas. Physical obstructions along the radar beam's path, such as mountains or buildings, can block or interfere with the radar signal, especially at lower elevation angles crucial for accurate radar-rainfall estimation. Other sources of uncertainty include the vertical variability of the precipitation system and the non-uniform vertical profile of reflectivity, which can complicate the conversion from measured reflectivity to precipitation depths. If there are significant amounts of snow and ice particles during winter at higher altitudes, stronger reflectivity signals in the form of noise are present as well. Additionally, stratiform and orographic rainfall can be underestimated due to overshooting, which is a concern in mountainous areas and at long ranges from the radar. Furthermore, there are range effects and temporal sampling errors that contribute to the uncertainty in radar measurements. These factors are compounded by issues like the polar-to-Cartesian grid transformation, variability in transmitted power, interference from wireless internet devices, non-uniform beam filling, and other technical challenges.

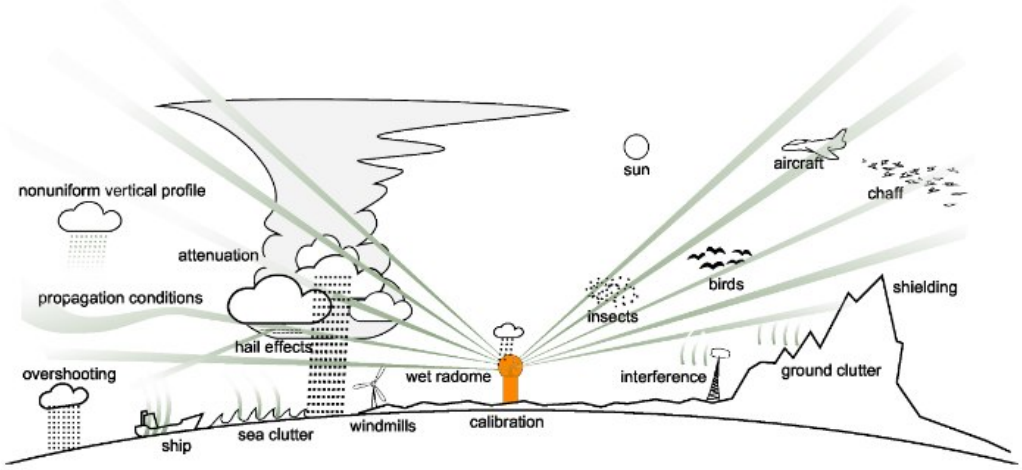


Figure 2.4: Phenomena that affect radar data quality (Holleman et al. 2005)

In order to account for these aforementioned errors the German Weather Service (Winterrath et al. 2017) has been developing a series of Quantitative Precipitation Estimates (QPEs). These products are derived from raw reflectivity radar data and lead to two final datasets called RADOLAN (Radar Online Aneichung/Radar Online Adjustment) and RADKLIM (Radarklimatologie/Radar Climatology). RADOLAN contains precipitation height measurements adjusted on a 900x900 km nationwide grid, with 1 km resolution. It contains data from June 2005 until today and it exists in hourly (RADOLAN RW) and 5-minute resolution (RADOLAN RY). In order to convert the radar data into precipitation heights, the local radar field is transformed to the nationwide grid and various correction methods are applied. Noise and clutter pixels are removed, shading and beam blockage is reduced, and smoothing through gradient filters takes place. The preprocessing further includes the development of a custom Z-R relation, which is used to carefully convert the reflectivities (Z) to rain rates (R). The most important step is the adjustment of the reflectivity data to the ground station measurements (the quantities measured at the rain gauges), by calculating the differences and factors between the ground stations and the associated pixels of the radar field and then interpolating these point measurements across a grid. At this point, Winterrath et al. make the assumption that the comparison between the station and radar-measured precipitation values at each location of the ground stations is representative for that particular location. For every ground station, they first calculate both the differences and the ratios (factors) between the precipitation values recorded at the ground stations and those estimated by the radar at the corresponding grid cells. The following formula is used for this calculation, with F_i representing the factor and D_i the difference for each station i (Eq. 2.5):

$$F_i = \frac{RR_{i,Ground}}{RR_{i,Radar}}$$
$$D_i = RR_{i,Ground} - RR_{i,Radar} \tag{2.5}$$

In addition to the direct station-to-radar comparison, the differences between the radar estimates and the ground station measurements for the eight surrounding grid cells are also determined. Out of this nine-cell grid, which includes the central cell and the eight surrounding cells, the one with the smallest absolute difference to the ground station measurement is selected. This process is designed to mitigate the error introduced by atmospheric drift, which can cause hydrometeors detected by radar at altitude to be displaced from their point of measurement by the time they reach the ground. Out of these 9 pixels, the one with the least absolute difference is chosen, as it is considered the most representative of the actual precipitation at that station's location. This is repeated for all the stations, and the values are then interpolated across the entire RADOLAN grid to calibrate the radar-based precipitation estimates. During interpolation, a defined radius of 40 km is used to weight the influence of each station's factors and differences on the surrounding grid cells The weighing function that is used is designed to prioritize proximity, thus ensuring the station's pixel value reflects nearby station values without over-smoothing the data.

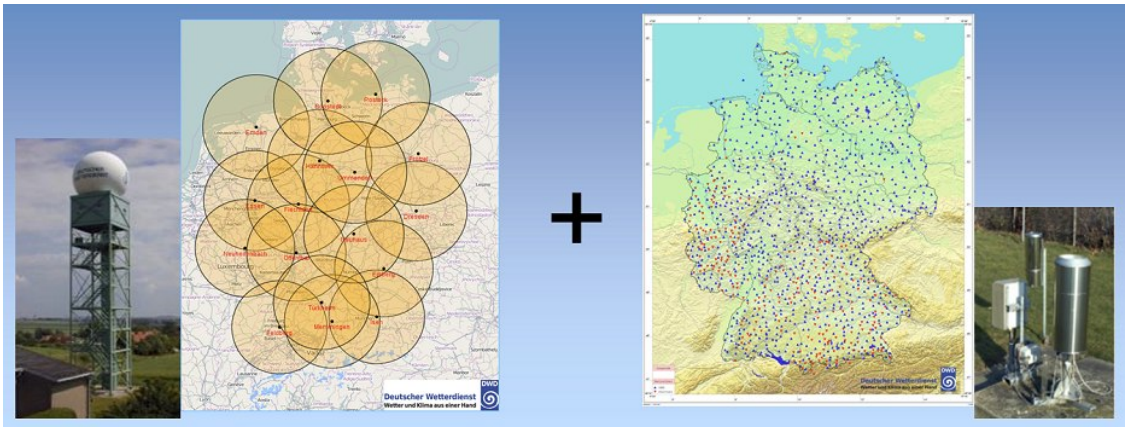


Figure 2.5: Combination of 17 radars and almost 1400 stations in the RADOLAN dataset (Source: RADOLAN - DWD)

Both the RADOLAN routine and the radar network have undergone continuous development, which improved the data quality over time. The first significant update in the processing routine occurred in December 2007, which included incorporating foreign and German gauges near the border outside any radar range and introduced gauge-based interpolation for gaps in the RW product. Subsequent major changes included extending the radar radius from 128 km to 150 km in March 2010 and the adoption of RY data for

generating the RW product from May 2010. This latter change added extra quality checks for clutter removal. In August 2016, an additional software update was implemented. This update introduced more clutter corrections, a filter to prevent biased rain gauge values from being used for adjustments, and reduced edges and inconsistencies at radar borders. More rain gauges from the Czech Republic were also considered for adjustments.

In 2018, the German Meteorological Service (DWD) released RADKLIM, a reanalyzed version of RADOLAN. This reanalysis utilized even more raw data processing techniques and incorporated more rain gauges for adjustment, marking a significant improvement in the quality and accuracy of the dataset. It includes several new climatological algorithms specifically designed to detect and correct radar-specific artifacts like clutter and spokes. Additionally, these algorithms address signal reduction with distance and height Other notable changes in RADKLIM are the reduction of the radius used for the adjustment of the radar data (in this case set to 128 km), and gaps in the radar data are no longer filled with interpolated gauge data. Instead, these gaps are represented in the dataset as missing values. The RADKLIM counterpart to the 5-minute RADOLAN RY product, known as RADKLIM YW, offers quasi-adjustment based on the RW product, thereby improving precipitation quantification. Moreover, the height of the produced map in RADKLIM has been extended by 200 rows (100 each in the north and south), and the grid has been shifted eastwards by 80 km. The resulting nationwide grid is 1100 x 900 km, compare to the square 900 x 900 km grid of RADOLAN. This expansion ensures that the dataset covers the entire country and provides additional zones on each side. The RADKLIM RW dataset, which is used in this thesis, contains hourly precipitation heights and currently spans from 2001-2020. An overview of RADOLAN and RADKLIM can be found in Table 2.2. In this work, the precipitation data that is used for the training of my model is part of the RADKLIM RW product.

Product	**Time Period**	**Temporal Resolution (min)**	**Radar Radius (km)**	**Grid (km)**
RADOLAN RW	June 2005 - present	60	125/150	900 x 900
RADOLAN RY	June 2005 - present	5	125/150	900 x 900
RADKLIM RW	2001 - 2020	60	128	1100 x 900
RADKLIM YW	2001 - 2020	5	128	1100 x 900

Table 2.2: Comparison of the RADOLAN and RADKLIM datasets

2.4 The Ground Station Network in Germany

The German Weather Service operates a dense network of ground stations scattered all across the country, measuring precipitation as well as other meteorological variables such as wind and temperature. They yield consistent measurements but are limited by their scattered distribution on the map, lacking the uniform field description characteristic of other methods. The longest observational time series goes back to the 18th century in Hohenpeißenberg, where temperature,pressure and precipitation have been measured since 1781. The majority of stations however started operating after 1945. More specifically, at the beginning of 1901 the precipitation network included 1400 stations, reaching a peak of 4500 stations in the 1980s. In 2012, the main (primary) network of the DWD, which is the source of data for weather reports, consisted of 507 stations measuring temperature and humidity, 212 pressure stations, 1925 precipitation stations, 297 wind stations and 302 stations measuring sunshine duration (Kaspar et al. 2013). The densest network is the one measuring precipitation, since this parameter exhibits the largest spatial variability.

Over time the network benefited from significant technological improvements. It coincided with the increase in the frequency of data collection, shifting from monthly to daily, hourly and even 5-minute intervals. In a weather station, instruments like rain gauges (also known as ombrometers) measure the amount of precipitation. A standard rain gauge consists of a funnel leading into a cylindrical container that collects the rainwater and measures its amount using a scale. The measurement is usually expressed in millimeters. Despite their simplicity, these instruments require regular maintenance and manual data recording, making them less practical for remote or inaccessible regions. Automatic rain gauges, on the other hand, are able to record data electronically and immediately deliver it to a database. The first electronic precipitation sensors were introduced in the late 1980s and early 1990s, but it was not until the beginning of 2000 that these new devices were incorporated in the network (Fig. 2.6). As a consequence of the automatization of the network, the number of stations decreased as they were now able to provide measurements at a higher temporal resolution (Fig. 2.6). In the beginning of 2000s, only around 100 out of more than 4000 precipitation stations were automated. Stations that measured in an hourly resolution rapidly started increasing only after 2004. As of the most recent update, the precipitation stations that provide hourly measurements are around 2000. However, only 200 of these stations comprise the main network of the DWD. They undergo quality control and are used to produce the synoptic reports. In this work I am using 76 stations that provide a long time series of precipitation measurements before the 2000s, at an hourly resolution (Table 2.3).

Employing these 76 station coordinates, I construct a binary station mask, similar to the missing value masks used in the original publication by Liu et al. More specifically, for each station's latitude and longitude value I determine the closest grid cell on the RADKLIM grid. This cell is assigned to the value of 1, signifying the existence of a station. All other cells in the grid are assigned to 0, indicating the absence of a station.

Consequently, this binary mask ensures that the model uses data from the grid cells flagged with a 1—those that represent the actual station locations. The model is thereby constrained to rely solely on the information from these marked cells and treats the rest of the cells as missing data points, similar to the approach of Liu et al. and Kadow et al. for image inpainting.

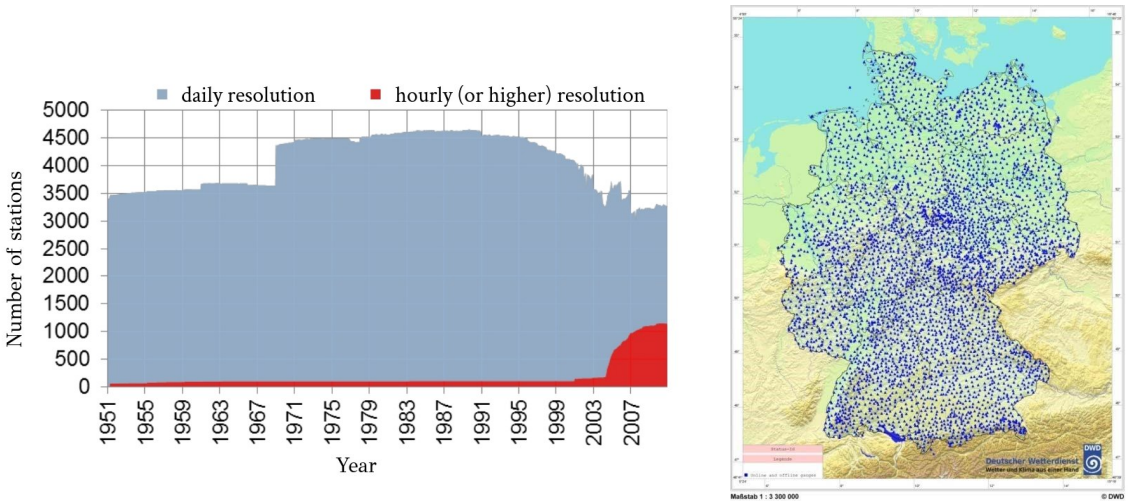


Figure 2.6: Number of precipitation stations and their respective temporal coverage through the years (left), modified from Müller et al. (2019). Map with the exact locations of more than 3000 stations in Germany (right), modified from Winterrath et al. (2017).

Station Name	Latitude	Longitude	Station Name	Latitude	Longitude	Station Name	Latitude	Longitude
Angermünde	53.03	13.99	Freudenstadt	48.45	8.41	Lüchow	52.97	11.14
Arkona	54.68	13.43	Gardelegen	52.51	11.39	Lüdenscheid	51.25	7.64
Artern	51.37	11.29	Garmisch-Partenkirchen	47.48	11.06	Magdeburg	52.10	11.58
Bamberg	49.87	10.92	Genthin	52.39	12.16	Mannheim	49.51	8.55
Barth	54.34	12.71	Gera-Leumnitz	50.88	12.13	Manschnow	52.55	14.55
Baruth	52.06	13.50	Görlitz	51.16	14.95	Marienberg	50.65	13.15
Berlin-Tegel	52.56	13.31	Göttingen	51.50	9.95	Meiningen	50.56	10.38
Berlin-Tempelhof	52.47	13.40	Goldberg	53.61	12.10	Michelstadt-Vielbrunn	49.72	9.10
Berus	49.26	6.69	Greifswald	54.10	13.41	Mühlacker	48.97	8.87
Boizenburg	53.39	10.69	Grünow	53.32	13.93	Mühldorf	48.28	12.50
Boltenhagen	54.00	11.19	Hamburg-Fuhlsbüttel	53.63	9.99	Marnitz	53.32	11.93
Braunlage	51.72	10.60	Hannover	52.46	9.68	München-Flughafen	48.35	11.81
Bremen	53.05	8.80	Harburg	48.79	10.71	Münster/ Osnabrück	52.13	7.70
Carlsfeld	50.43	12.61	Harzgerode	51.65	11.14	Neuruppin	52.90	12.81
Chemnitz	50.79	12.87	Hohenpeißenberg	47.80	11.01	Nürburg-Barweiler	50.36	6.87
Chieming	47.88	12.54	Kempten	47.72	10.33	Norderney	53.71	7.15
Cottbus	51.78	14.32	Klippeneck	48.11	8.75			
Doberlug-Kirchhain	51.65	13.57	Köln-Bonn	50.86	7.16	Oschatz	51.30	13.09
Dresden-Klotzsche	51.13	13.75	Konstanz	47.68	9.19	Osterfeld	51.09	11.93
Düsseldorf	51.30	6.77	Kyritz	52.94	12.41	Oberstdorf	47.40	10.28
Elpersbüttel	54.07	9.01	Lautertal-Oberlauter	50.31	10.97	Öhringen	49.21	9.52
Erfurt-Weimar	50.98	10.96	Leinefelde	51.39	10.31	Plauen	50.48	12.13
Essen-Bredeney	51.40	6.97	Leipzig/Halle	51.43	12.24	Potsdam	52.38	13.06
Fichtelberg	50.43	12.95	Lichtenhain-Mittelndorf	50.94	14.21	Pelzerhaken	54.09	10.88
Frankfurt/ Main	50.03	8.52	Lindenberg	52.21	14.12	Putbus	54.36	13.48
Freiburg	48.02	7.83	Lingen	52.52	7.31			

Table 2.3: List of the stations used to construct the binary station mask. They provide hourly measurements of precipitation since 1995 (Müller et al. 2019).

2.4 Model Description

In this thesis I use the original baseline model, as proposed by Liu et al. The network is based on the U-Net architecture (Ronneberger et al. 2015) but, as mentioned in Section 2.2, all convolutional layers are replaced with partial convolutions. The U-Net architecture is designed to capture both local and global contextual information in an image. It is named U-Net because of its U-shaped architecture that consists of an encoder path and a decoder path. Here, the encoder path consists of partial convolutions accompanied with pooling layers. The partial convolution layers apply filters to the input data in order to extract features, while pooling layers reduce the spatial dimensions of the data, thus increasing the depth of the network. On the other hand, the decoder is symmetric to the encoder, but the max pooling layers are replaced by upsampling layers while the partial convolutions remain as is. The upsampling is performed using nearest neighbour interpolation and allows to gradually increase the spatial resolution of the data until it matches the original input size.

Another important characteristic of the U-net is its skip connections, that serve as the bridge of communications between layers of equal resolution in the encoder and decoder. They concatenate feature maps from the encoder with the corresponding feature maps in the decoder, allowing the network to retain detailed information from the early layers while keeping high-level contextual information. In the context of climate sciene, high-level contextual information refers to capturing the larger climate patterns and relationships between different variables or regions. In Ronneberger's original work where the U-Net was first introduced, skip connections are crucial for precise localization in image segmentation, as they help recover the spatial information lost during downsampling. In the work by Liu et al., these skip connections play a similar role but are more focused on ensuring that the infilled areas are consisted with the rest of the image.

A crucial component that I adapt from the original publication is a specialized loss function with various terms, each one targeting different aspects of the infilling procedure and its evaluation. In principle, a loss function measures the similarity between the model's prediction and the ground truth. During training, the model adjusts its parameters to minimize the loss function and improve its predictions. The first two terms of the loss function I use here (L_{valid} and L_{hole}) are the L_1 losses, or the Mean Absolute Error between the true $I_{ground\,truth}$ grid and the predicted I_{output} grid, for valid and missing (hole) pixels (Eq. 2.6a and 2.6b). These terms control the pixel-wise reconstruction accuracy of the model. In the following equations, M denotes the mask, $I_{ground\,truth}$ is the

ground truth image, I_{output} is the model output, and N is the total number of values (grid cells) in each input grid.

$$L_{hole} = \frac{1}{N_{I_{ground\,truth}}} L_1\left((1-M)\odot\left(I_{out} - I_{ground\,truth}\right)\right) \tag{2.6a}$$

$$L_{valid} = \frac{1}{N_{I_{ground\,truth}}} L_1\left(M\odot\left(I_{out} - I_{ground\,truth}\right)\right) \tag{2.6b}$$

Next comes the perceptual loss term (L_{prc}) which is designed to capture and minimize the high-level perceptual and semantic differences between output and ground truth images, going beyond mere pixel-level accuracy (Gatys et al. 2015). It operates by utilizing a pre-trained neural network, the VGG16, which has already been trained to classify images from the ImageNet dataset into 1000 classes (Simonyan et al. 2014). The pre-trained network extracts feature maps from the ground truth $I_{ground\,truth}$, the output I_{output} , as well as the output composition I_{comp} (the combination of the raw model output image with the valid pixels directly taken from the ground truth image. Then, the differences between these feature representations are calculated, enabling the capture of low and high level features in the data (Eq. 2.7). In the equation below, p denotes the number of layers, N is again the number of elements and Ψ are the feature maps.

$$L_{prc} = \sum_{p=0}^{P-1} L_1 \frac{\left(\Psi_p^{I_{out}} - \Psi_p^{I_{ground\,truth}}\right)}{N_{\Psi_p^{I_{groundtruth}}}} + \sum_{p=0}^{P-1} L_1 \frac{\left(\Psi_p^{I_{comp}} - \Psi_p^{I_{ground\,truth}}\right)}{N_{\Psi_p^{I_{groundtruth}}}} \tag{2.7}$$

The fourth term of the loss function is the style loss L_{style} which measures the differences between the Gram Matrices of the prediction and of the Ground Truth. In computer science, the Gram matrix is a representation of the correlation between the feature maps of a certain layer in the network. For an image processed through the network, the Gram matrix is calculated by multiplying the feature map matrix with its transpose. The Gram matrix represents the correlations between different feature maps and it contains the style information of an image such as textures and colors. Two style losses are calculated: one for the direct model output I and one for the output composition I_{comp} (Eq. 2.8a and 2.8b). In the equations below, C_p is the channel size, H_p is the image height and W_p is the image width at layer p:

$$L_{stlye,out} = \sum_{p=0}^{P-1} \frac{1}{C_p C_p} L_1\left(\frac{1}{C_p H_p W_p}\left(\left(\Psi_p^{I_{out}}\right)^T\left(\Psi_p^{I_{out}}\right) - \left(\Psi_p^{I_{groundtruth}}\right)^T\left(\Psi_p^{I_{ground\ truth}}\right)\right)\right) \quad (2.8a)$$

$$L_{stlye,comp} = \sum_{p=0}^{P-1} \frac{1}{C_p C_p} L_1\left(\frac{1}{C_p H_p W_p}\left(\left(\Psi_p^{I_{comp}}\right)^T\left(\Psi_p^{I_{comp}}\right) - \left(\Psi_p^{I_{groundtruth}}\right)^T\left(\Psi_p^{I_{ground\ truth}}\right)\right)\right) \quad (2.8b)$$

The final term is the total variation loss, which is a common loss function in image processing tasks. It is based on the concept of total variation, which measures the variations or changes in pixels with respect to their neighbors. The loss term computes how strong the smoothing across pixels within a 1-pixel dilation (i,j+1) from the hole is (Eq. 2.9).

$$L_{tv} = \sum_{(i,j)\in R,(i,j+1)\in R} L_1 \frac{\left(I_{comp}^{i,j+1} - I_{comp}^{i,j}\right)}{N_{I_{comp}}} + \sum_{(i,j)\in R,(i+1,j)\in R} L_1 \frac{\left(I_{comp}^{i+1,j} - I_{comp}^{i,j}\right)}{N_{I_{comp}}} \quad (2.9)$$

All the separate loss terms are combined to a total loss Eq. 2.10) to quantify the difference between the model's prediction and the actual ground truth. This is a measure of how well or poorly the model is performing on the given task. The weights in the following equation were adapted from the original publication by Liu et al., and according to the authors they were determined based on a hyperparameter search on a validation set of 100 images.

$$L_{total} = L_{valid} + 6L_{hole} + 0.05L_{prc} + 120\left(L_{style_{out}} + L_{style_{comp}}\right) + 0.1L_{tv} \quad (2.10)$$

The calculated loss is used to perform backpropagation. This involves calculating the gradients of the loss with respect to the model parameters (weights). These gradients represent the direction and magnitude of the change needed to minimize the loss. The model weights are updated using optimization algorithms like gradient descent. The purpose is to adjust the model parameters in a way that reduces the loss, leading the model to perform better. The entire process (forward pass, loss calculation, backpropagation, and weight update) constitutes one training cycle. This cycle is repeated multiple times (epochs) to iteratively improve the model's performance.

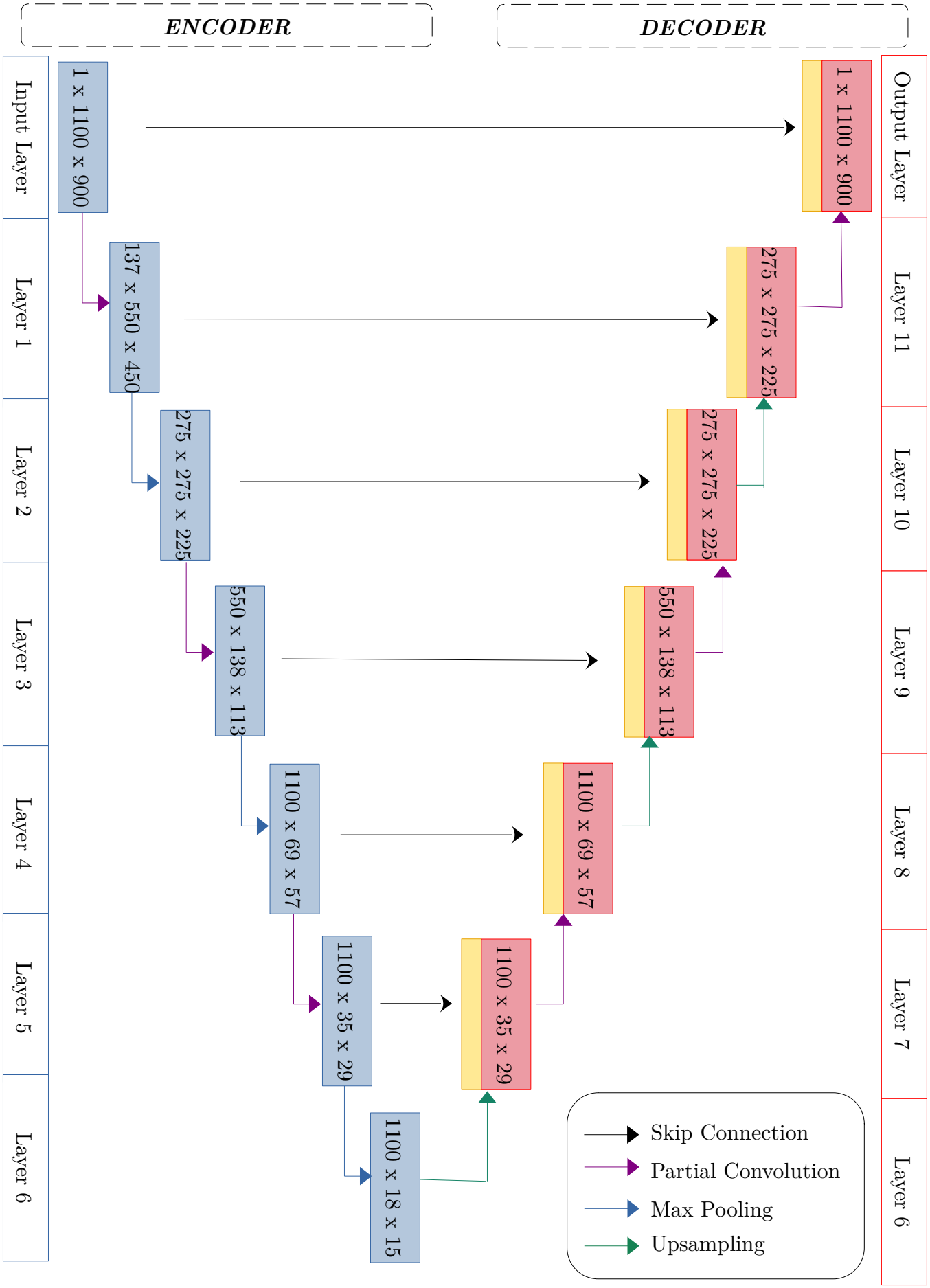


Figure 2.7: Graph of the model's U-Net architecture. The yellow boxes indicate the information passed from the encoder to the decoder through the skip connections (black arrows).

2.5 Training Configuration

At the start of each training iteration, the data loader provides samples with a batch size of 4. Each data sample consists of meteorological variables measured at the stations, the binary station mask M and the ground truth RADKLIM precipitation grid G. The samples are received by the model as tensors with dimensions C x H x W. C stands for the number of channels and corresponds to the number of input observational meteorological variables, while H and W are the grid height and width correspondingly. By applying the binary station mask remapped on the RADKLIM precipitation grid, the input image that the model receives has a total of 76 valid points, placed at the station locations. The task of the model is to use these 76 valid points to reconstruct a full precipitation field. The reconstructed image is compared to the corresponding complete RADKLIM grid (the ground truth), and the training losses are calculated. As explained in Section 2.4, models are optimized by minimizing the training loss iteratively until a certain number of iterations. The analysis of the learning curves (training and validation losses as a function of the number of iterations) provides evidence of the model's fitness and the absence of overfitting.. The training process is depicted in Figure 2.8.

An overview of all the experiments is given in Table 2.4. For the training periods listed, there were certain timesteps where RADKLIM grids were incomplete due to radar failures. For consistency purposes, these timesteps were excluded from training, and thus only full radar grids were considered. The most straightforward method was to solely use precipitation measurements as input and have the model reconstruct RADKLIM grids as output. I performed 2 runs of this type, run #1 and run #8, with the only difference being the size of the training set. Apart from precipitation, I used pressure (measured in Pascal), surface wind speed (measured in m/s) and surface temperature (measured in Kelvin), incorporating them in runs 3-7 and 10. Due to memory issues related to the inclusion of all meteorological variables, run #4 presents the shortest time span for the training years. When two meteorological variables were used per run, I was able to train with the 2001 – 2006 setup. I also made an experiment where instead of providing the model with absolute values of pressure, I subtracted the pressure measurements for consecutive timesteps. This approach includes the information of pressure change within an hourly time window, and this temporal aspect can be valuable for the model to capture dynamic patterns and trends in the data. Moreover, the subtraction step greatly reduces the magnitude of the values: the absolute pressure values ranged from 900-1040 hPascal, while the differences were within the range of 0.5-2 hPascal. Smaller input values can be very advantageous, as they may lead to more stable training and potentially reduce issues related to numerical precision. They can also decrease the impact of outliers or extreme values. Another approach was to consider more than one timestep at each iteration (runs 2 and 9): to predict a single timestep, I use 2 timesteps before and 2 timesteps after the target timestep. Finally, in the most advanced approach I incorporated both the multiple timestep methodology, as well as the pressure change, in run #10.

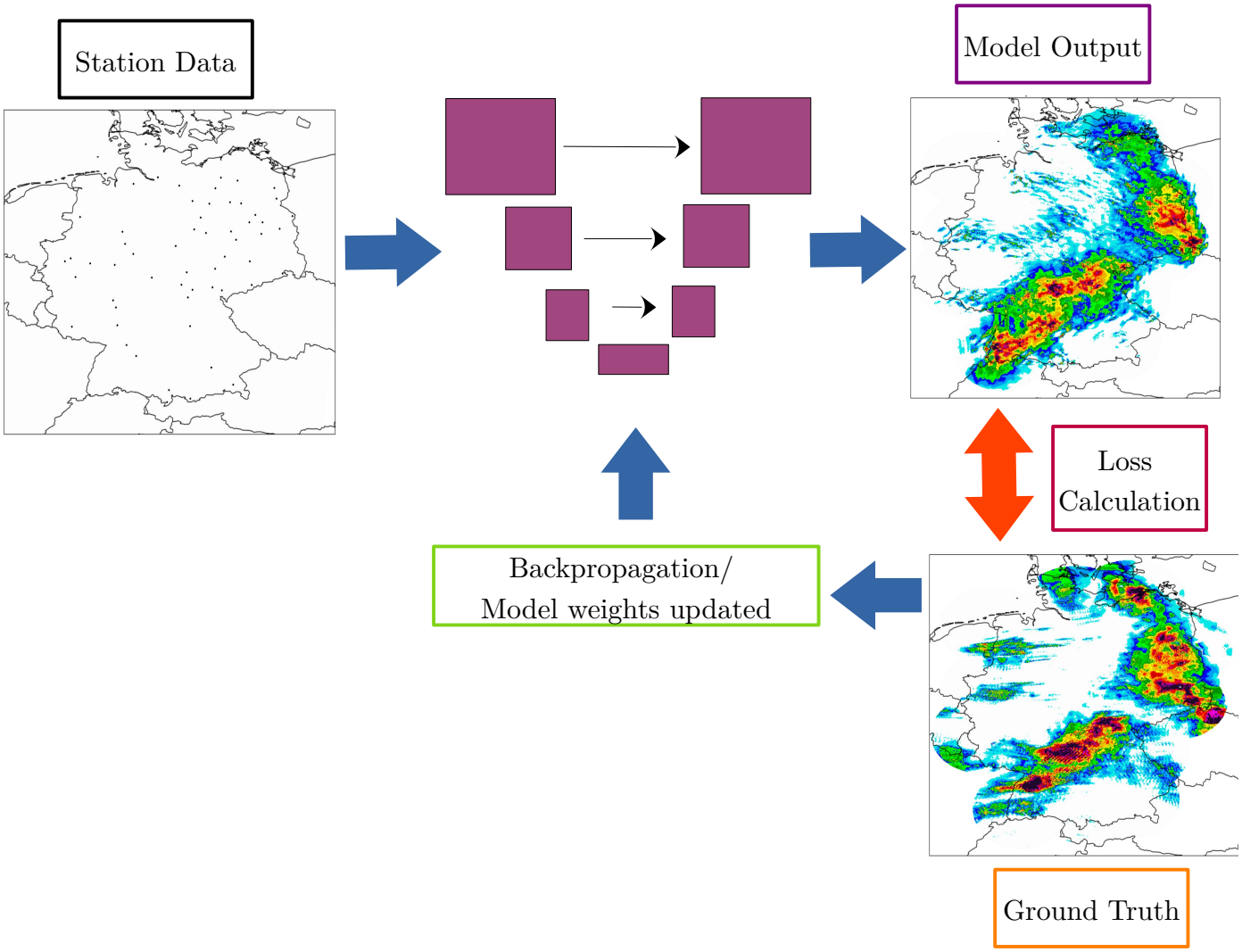


Figure 2.8: Overview of the training process. The input image (left) is loaded into the model (middle) and an output is generated. The output is compared with the original RADKLIM grid which corresponds to the Ground Truth and the loss is calculated via the loss function 2.10. Then the backpropagation process begins and the model weights are updated, as a new training cycle begins.

All the runs were trained for 80,000 iterations, but due to the differences in training dataset size, the real-time training varied from 1 day to 4 days. The trained models were used to reconstruct the years 2018 and 1997, and evaluation was performed on the reconstructions, with a series of metrics outlined in the next chapter. The learning rate was set to 0.0002 for all runs. Backpropagation utilized the Adam optimizer (Kingma et al. 2014), an enhanced weight update method in contrast to the traditional stochastic gradient descent, for more efficient model parameter adjustments.

Run Number	Training Years	Validation Years	Test Years	Input Station Data	Output Data
1	2001-2011	2012-2017	1997,2018	precipitation	RADKLIM
2	2001-2011	2012-2017	1997,2018	precipitation, timesteps 2+2	RADKLIM
3	2001-2006	2007	1997,2018	precipitation, pressure change	RADKLIM
4	2001-2004	2005	1997,2018	precipitation,wind,pressure, temperature	RADKLIM
5	2001-2006	2007	1997,2018	precipitation, wind	RADKLIM
6	2001-2006	2007	1997,2018	precipitation,temperature	RADKLIM
7	2001-2006	2007	1997,2018	precipitation, pressure	RADKLIM
8	2001-2006	2007	1997,2018	precipitation	RADKLIM
9	2001-2006	2007	1997,2018	precipitation,timesteps 2+2	RADKLIM
10	2001-2006	2007	1997,2018	precipitation, pressure change, timesteps 2+2	RADKLIM

Table 2.4: Overview of the experiments

2.6 Reference Datasets

Throughout this thesis, I reference several gridded datasets. I use the HYRAS-PRE dataset to calculate the performance evaluation metrics of the models in Chapter 3. Given its daily resolution, HYRAS was not the primary choice for investigating specific events in Chapter 4. Instead, I opted for two reanalysis datasets, ERA5 and REA2, in that chapter. This choice allowed for a comparative analysis of my model against the reanalysis methodology.

The HYRAS dataset (Rauthe et al. (2013)) is a comprehensive collection of high-resolution gridded climate data predominantly focused on Germany and its surrounding regions. The HYRAS-PRE dataset is the precipitation component, it uses data from 6200 stations across various countries and it covers the years from 1951 to the present on a daily temporal resolution, at 1 km^2 grid scale. It includes all the river basins of Germany and also data from neighboring countries that share the same river basins. Thus, data from Netherlands, Belgium, Luxembourg, France, Austria, Czech Republic and Switzerland are

also incorporated into the dataset. The interpolation method that is used for the creation of HYRAS-PRE is called REGNIE (Regionalisierte NIEderschlagshöhe/Regionalised Precipitation Amount). It is a combination of multiple linear regression (MLR) and inverse distance weighting. It consists of two main steps: first, mean background precipitation fields are calculated, and secondly the daily data are interpolated as a ratio of the total precipitation to the climatology. This methodology has been traditionally used for years at the DWD and its main advantage is that it incorporates the original measured precipitation amounts at the stations, without smoothing them out.

Even though HYRAS is a lengthy and heavily consistent daily time series, it has several drawbacks. Station network density varied over the years that the dataset was created for. As stated by the authors, this variability can negatively affect the quality of the dataset. Another problem is related to the different measuring types (due to different rain gauge models that were used) and various measuring timestamps (e.g. 8:00 UTC at Netherlands, 6:30 UTC at Czech Republic, 6:00 UTC in the rest of the countries). These inconsistencies were neglected in the creation of the HYRAS dataset. Furthermore, precipitation measurements often suffer from a bias in accuracy. More specifically, it is common that the true precipitation amount is underestimated by at least 10%, as Rauthe et al. (2013) state. This bias was not accounted for.

ERA5 (Hersbach et al. 2020) is the fifth atmospheric global reanalysis dataset produced by the European Center for Medium-Range Weather Forecasts (ECMWF), replacing its predecessor, ERA-Interim, which was released in 2006. ERA5 uses a 4D-Var data assimilation system with a modern global atmospheric model, the ECMWF Integrated Forecasting System (IFS), to include measurements from different observation sources (stations, satellites etc.). This assimilation process, conducted incrementally over 12-hour windows at a spatial resolution of 31 km, weaves together various data sources to produce a comprehensive and physically coherent analysis across atmosphere, ocean, and land. The dataset covers the global climate from January 1940 to the present in an hourly resolution, while the atmospheric resolution extends through 137 levels up to 80 km in height. The precipitation forecasts in ERA5 benefit from a sophisticated precipitation scheme, informed by microphysics, cloud dynamics, and convection processes, resulting in marked improvements in precipitation representation.

Finally, the REA2 dataset, or COSMO-REA2 (Wahl et al. 2017), is a convective-scale regional reanalysis system for Central Europe, developed by the Hans-Ertel Centre for Weather Research - Climate Monitoring Branch. It is based on the COSMO (Consortium for Small-Scale MOdeling) model and employs a data assimilation technique that continuously nudges various observational data including radar scans, aircraft measurements, wind profilers, and station data. It also uses latent heat nudging of radar-derived precipitation every 5 minutes. The reanalysis covers Central Europe (Austria, Belgium, Denmark, Germany, Liechtenstein, Luxembourg, The Netherlands, Slovenia and Switzerland, as well as parts of the Czech Republic, France, Italy, Poland and the UK)

with a spatial resolution of 0.018° (2 km) grid spacing, encompassing 724x780 grid points across 50 vertical levels. The time range of the dataset spans from 2007 to 2013 with an hourly temporal resolution. REA2 is particularly notable for its representation of precipitation. High-resolution reanalyses like REA2 can improve the representation of precipitation over Europe compared to coarser gridded global reanalyses. However, there are challenges due to the discrepancy between gridded precipitation model data and localized observations, especially in the representation of extreme events. While increasing horizontal resolution improves the representation of extremes, it also introduces the possibility of displacement errors (Wahl et al. 2017).

3. Results

In this Chapter I present the results of my work. I implement a range of metrics and performance scores. Specifically, I focus on using the trained models to reconstruct the precipitation fields for the years 2018 and 1997. The selection of these years is not random: 1997 falls within the period of 1995-2000, which is my target for accurate reconstruction. Additionally, I chose the year 2018 since it falls beyond the 2011-2017 range to ensure my evaluation is conducted on data not previously used in the training or validation sets.

For 2018, the model reconstructions are compared against two reference datasets: RADKLIM and HYRAS (Rauthe et al. 2013). Consequently, all chosen metrics – including the root mean square error (RMSE), time and field correlation, and the difference of total precipitation fall – are calculated by comparing my reconstructed 2018 data with the corresponding fields from these reference datasets. For the evaluation of the year 1997, I encounter a limitation due to the unavailability of the RADKLIM dataset for the years before 2001. As a result, my performance analysis for 1997 is exclusively based on comparisons with the HYRAS-PRE dataset. The aforementioned metrics in this case are thus calculated solely against the HYRAS-PRE data for the 1997 reconstructions.

3.1 Investigation of Model Prediction Accuracy through the RMSE

My analysis begins by employing two distinct methodologies to calculate the following variations of the RMSE: the monthly mean of the field averaged RMSE (Eq. 3.1), which is represented by a time series (Fig. 3.1), and the RMSE averaged both in space and time (Eq. 3.2), which is a single value. I choose these variations since precipitation is highly variable in space and time, and thus it is important to investigate it through multi-dimensional metrics and not solely through single values. In the formulas below, R is the

reference dataset (HYRAS or RADKLIM), O is the model output, N is the number of timesteps, W and H are the grid width and height respectively. Figures 3.1-3.3 depict the monthly RMSEs derived from Eq. 3.1 while Table 3.1 shows all the RMSE scores calculated using Eq. 3.2.

$$RMSE = \sqrt{\frac{1}{W\,H}\sum_{i=0}^{W-1}\sum_{j=0}^{H-1}\left(R_{t,i,j} - O_{t,i,j}\right)^2} \qquad \text{(Eq. 3.1)}$$

$$RMSE = \sqrt{\frac{1}{N\,W\,H}\sum_{t=0}^{N-1}\sum_{i=0}^{W-1}\sum_{j=0}^{H-1}\left(R_{t,i,j} - O_{t,i,j}\right)^2} \qquad \text{(Eq. 3.2)}$$

In Table 3.1 the last row (RADKLIM), calculated only for the first column, represents the averaged RMSE between HYRAS and RADKLIM for 2018. It is calculated after averaging the original hourly RADKLIM at a daily resolution (to match HYRAS). I calculate this quantity to have a further reference to compare the models against. It is important to note that HYRAS and the models' output have a different grid so the outputs were remapped into the HYRAS grid using Nearest Neighbor remapping. I did the same for the RADKLIM-HYRAS reference. I expect that this step will introduce some error in my calculations, which was not accounted for.

Run	**RMSE with HYRAS 2018 (mm/h)**	**RMSE with RADKLIM 2018 (mm/h)**	**RMSE with HYRAS 1997 (mm/h)**
1	**0.1484**	**0.1611**	**0.1630**
2	0.1565	0.1642	0.1683
3	0.1529	0.1646	0.1657
4	0.1630	0.1731	0.1793
5	**0.1474**	**0.1594**	**0.1620**
6	0.1508	0.1669	0.1696
7	0.1518	0.1648	0.1673
8	**0.1499**	0.1632	**0.1649**
9	0.1532	**0.1625**	0.1655
10	0.1605	0.1681	0.1721
RADKLIM	0.1093	-	-

Table 3.1: All the single-value RMSEs calculated for the reconstructed years 1997 and 2018, with respect to the reference datasets RADKLIM and HYRAS-PRE. The bold numbers highlight the top three best performing models in each column.

I see from the Table 3.1 that for the year 2018 the HYRAS-RADKLIM pair yields the lowest RMSE value. Given that my training dataset predominantly consists of RADKLIM data, it is unrealistic to expect my models to surpass this benchmark. Nonetheless, this reference RMSE serves as a valuable target, providing a standard against which to measure the models' performance. Essentially, the closer a model's RMSE is to that of the HYRAS-RADKLIM reference, the more accurate the model is considered to be. I begin by examining the RMSE scores in the first column, which are calculated based on the comparison between my models' outputs and the HYRAS-PRE data for the year 2018. The models achieving the lowest RMSE scores (below 0.15 mm/h) are runs 5, 1, and 8. Specifically, run 5 incorporates station wind measurements, which appear to enhance the model's performance. In contrast, runs 1 and 8 achieve similar accuracy without the integration of additional information beyond precipitation. In the second column, I observe that runs 1 and 5 continue to perform well, accompanied by run 9. This latter run, which exclusively utilizes precipitation data, also benefits from temporal information as it employs multiple timesteps for each prediction. This approach's effectiveness agrees with previous studies (Meuer et al. 2022), and it highlights the value of temporal information in precipitation studies. In the third column where the results from the year 1997 are laid out, the same runs - 5, 1, and 8 – once again emerge as the best. Run 4, which integrates a comprehensive set of station measurements including wind, temperature, pressure, and precipitation, records the highest RMSE score across all three evaluated scenarios. Similarly, Run 10, despite incorporating temporal information and pressure measurements, fails to deliver a strong performance. These results suggest that the inclusion of multiple variables does not necessarily guarantee improved model performance, or perhaps the new variables need to be incorporated with a more sophisticated approach.

It is important to note that the gap in performance between the top runs and the others is relatively narrow, with a difference of merely around 0.02 mm/h. This observation suggests that while certain runs exhibit marginally superior RMSE scores, the overall model performance across different runs is quite comparable. To gain a comprehensive understanding of all models' performance, it is essential to consider additional metrics beyond the single-value RMSE. This includes examining time series and spatial distribution maps, which will allow us to uncover potential regional or seasonal variations in model performance, offering a more comprehensive view of their effectiveness.

Figure 3.1 depicts the monthly mean RMSE for the year 2018 calculated between my models' output and HYRAS. The dashed black line is the monthly mean RMSE

Run configurations: 1 (precipitation, long train), 2 (precipitation, 2 timesteps, long train), 3 (precipitation, pressure change), 4 (precipitation, pressure, temperature, wind), 5 (precipitation, wind), 6 (precipitation, temperature), 7 (precipitation, pressure) ,8 (precipitation, short train), 9 (precipitation, 2 timesteps, short train) , 10 (precipitation, pressure, 2 timesteps) (see also Table 2.4)

calculated between HYRAS and RADKLIM, to be used as a reference in a similar manner with Table 3.1. This reference pair exhibits once again the lowest RMSE value, remaining constantly below all the model runs with significant difference. Figure 3.2 shows the same monthly trends, but calculated using RADKLIM as the reference dataset.

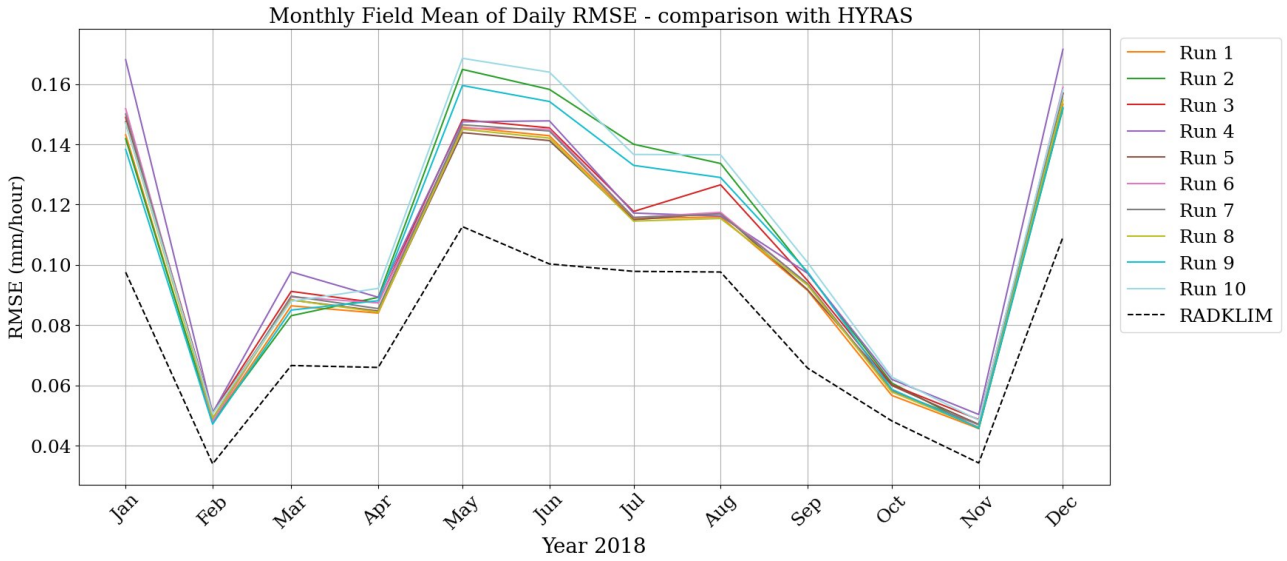


Figure 3.1: Monthly field mean of daily RMSE at each grid point of the year 2018 calculated using HYRAS as reference. Colored lines represent the model runs while the black dashed line represents the RMSE calculated between HYRAS and RADKLIM.

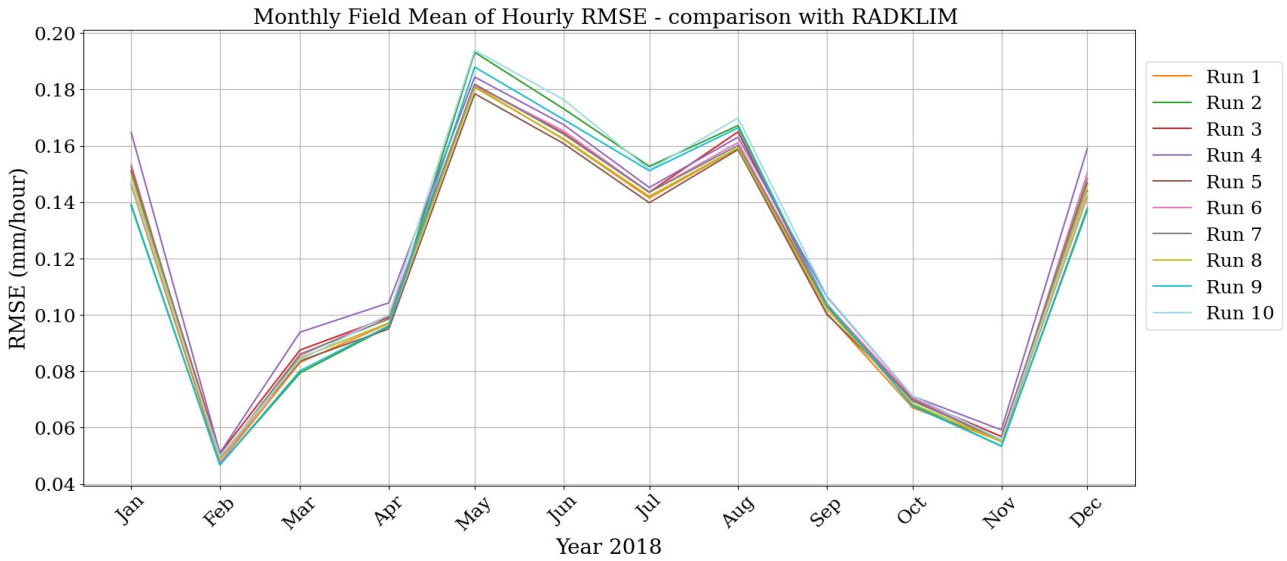


Figure 3.2: Monthly field mean of hourly RMSE at each grid point of the year 2018 calculated using RADKLIM as reference. Colored lines represent the model runs.

Upon investigating monthly RMSE plots, distinct patterns emerge throughout the year. In the initial months (January, February, March) and the concluding months

(November, December), run 4 consistently exhibits the highest error. In the middle of the year, the error for run 4 falls close to the other runs. This pattern is consistent in the comparisons with both HYRAS and RADKLIM datasets. Run 3, which integrates precipitation data and the change in pressure over time, displays a notable decrease in performance in August when compared to other runs, yet closely matches their error magnitude for the remainder of the months. Another notable trend is observed from May to August, where runs 2, 9, and 10 show the highest errors. Despite this, runs 2 and 9 record the smallest errors for the rest of the year. The similarity in their performance can be attributed to their shared training methodology, which involves multiple input timesteps (two past and two future). Their difference is the training dataset size; run 9 is trained on nearly half the data compared to run 2. Still, run 9 achieves a lower error than run 2, as seen also in the RMSE values presented in Table 3.1. Run 1, ranking as the second-best model in Table 3.1, consistently maintains the lowest error from May to September. Run 5, identified as the top performer in Table 3.1, showcases one of the lowest errors from April to August. However, it does not particularly stand out as either one of the best or worst models for the rest of the year.

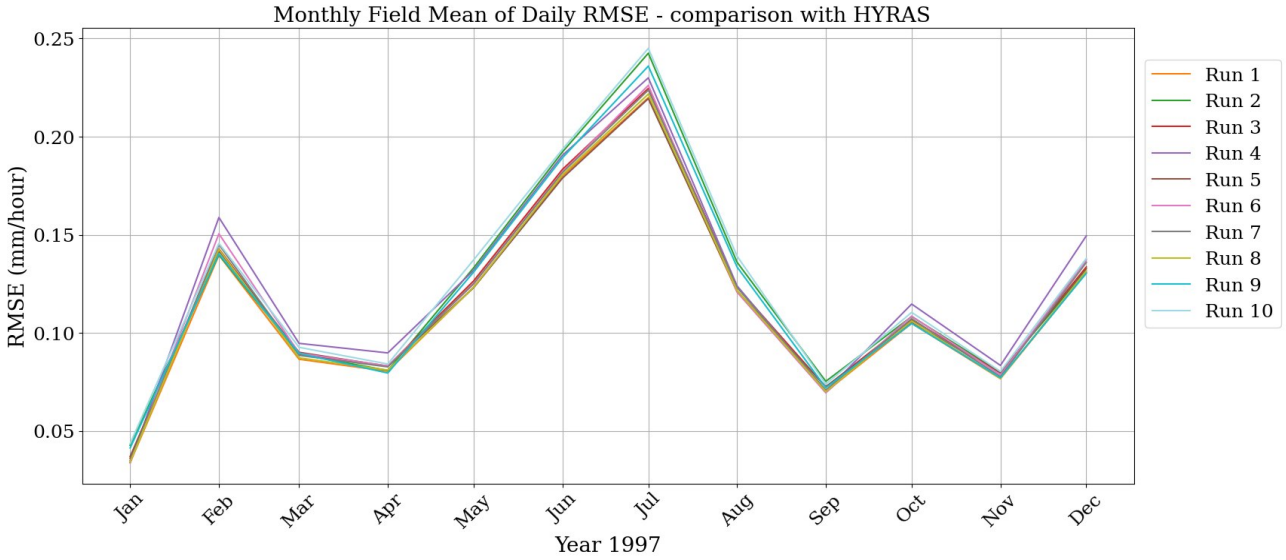


Figure 3.3: Monthly field mean of daily RMSE at each grid point of the year 1997 calculated using HYRAS as reference. Colored lines represent the model runs

In the 1997 plots, run 5 demonstrates the lowest error exclusively in July; however, for the remaining months, its error is comparable with the other runs. Run 4, matches its 2018 performance, and it has one of the poorest performances, particularly from January to April and from October to December. In contrast, runs 1, 3, 7, and 8 do not exhibit any notably distinctive performance trends. Compared to 2018, runs 9 and 2 maintain similar error profiles and both reach their peak in July. Interestingly, they record the lowest errors during February, April, and the final quarter of the year – October to December. Run 10,

while displaying the highest error in July when compared to the rest, holds unremarkable RMSE values for the rest of the year. Notably, run 6 presents a higher error in 1997 than in 2018, indicating a deviation in its yearly performance.

The analysis of Figures 3.1, 3.2, and 3.3 reveals no clear-cut relationship in terms of monthly predictive skill across the models. Some similarities in error magnitude are observed in March, April, and August for both years under study, yet no further patterns emerge. To gain deeper insights into the models' performance and their month-specific predictive abilities, I calculate the field correlations between the models' outputs and the reference datasets.

3.2 Spatial Correlation between Model and Reference Data

The Pearson Correlation Coefficient, often encountered in meteorological and climatological studies, is a statistical tool used to measure the degree of linear relationship or correlation between two spatial fields. Here, I will use this coefficient to evaluate how well my model's output and the reference datasets agree with each other in terms of spatial distribution and patterns. In essence, the Pearson Correlation Coefficient, when applied to field data, compares two gridded datasets on a point-by-point basis. A high correlation coefficient indicates that the spatial patterns in the two datasets are very similar, suggesting that the model is accurately capturing the spatial distribution of my two reference datasets, RADKLIM and HYRAS.

The coefficient value ranges from -1 to +1. A value of +1 implies a perfect positive correlation, meaning that as values in one dataset increase, values in the corresponding locations of the other dataset increase proportionally. A value of -1 indicates a perfect negative correlation, where an increase in values in one dataset corresponds to a proportional decrease in the other. A value of 0 signifies no linear correlation between the datasets.

First, I calculate the field correlation coefficient according to Eq. 3.3. This formula calculates the Pearson Correlation Coefficient across corresponding spatial fields of two datasets. This process allows us to evaluate the degree to which my model's output and the reference datasets exhibit similar spatial patterns, or how well my models capture the spatial distribution of precipitation. A high correlation coefficient (close to 1) in a precipitation model suggests that the model is successful not only in predicting the right amount of rainfall but also in accurately forecasting where this rainfall occurs. A model that predicts the right amount of rainfall but misplaces its geographical location has a low field correlation, and would still require improvement.

$$P_{field\ cor} = \frac{\sum_{t=0}^{N-1} R_{t,i,j} O_{t,i,j} - \overline{R_{i,j}}\,\overline{O_{i,j}}}{\sqrt{\left(\sum_{t=0}^{N-1} R_{t,i,j}^2 - \overline{R_{i,j}}\right)\left(\sum_{t=0}^{N-1} O_{t,i,j}^2 - \overline{O_{i,j}}\right)}} \qquad \text{(Eq. 3.3)}$$

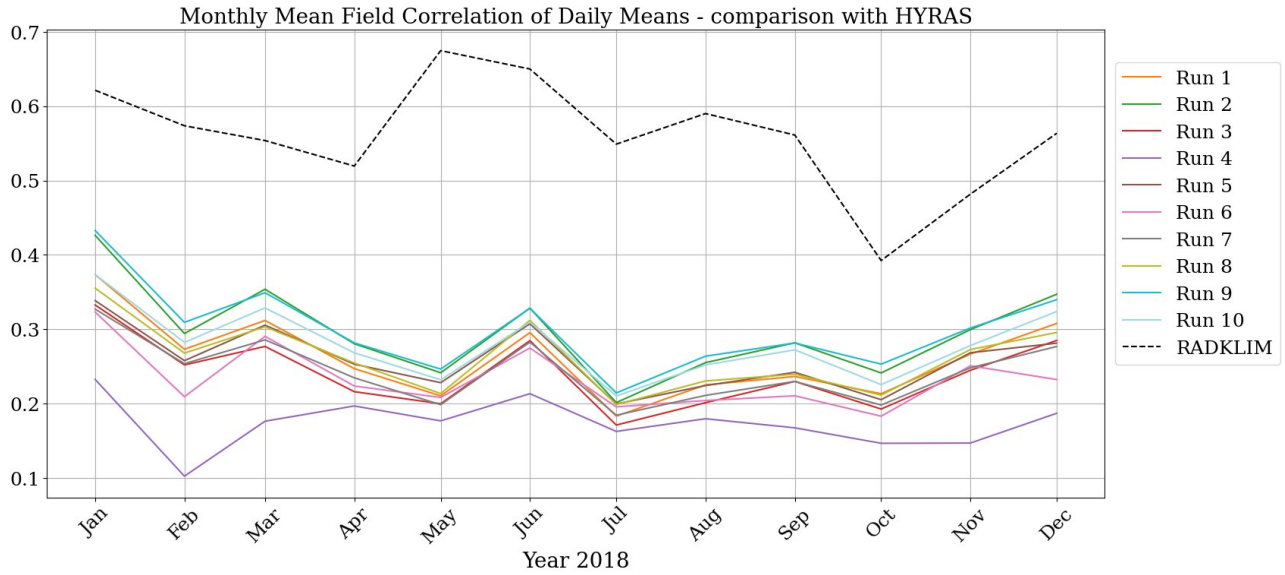


Figure 3.4: Monthly mean Field Correlation of the year 2018 calculated at each grid point using HYRAS as reference. Colored lines represent the model runs while the black dashed line represents the Field Correlation calculated between HYRAS and RADKLIM.

Run configurations: 1 (precipitation, long train), 2 (precipitation, 2 timesteps, long train), 3 (precipitation, pressure change), 4 (precipitation, pressure, temperature, wind), 5 (precipitation, wind), 6 (precipitation, temperature), 7 (precipitation, pressure) ,8 (precipitation, short train), 9 (precipitation, 2 timesteps, short train) , 10 (precipitation, pressure, 2 timesteps) (see also Table 2.4)

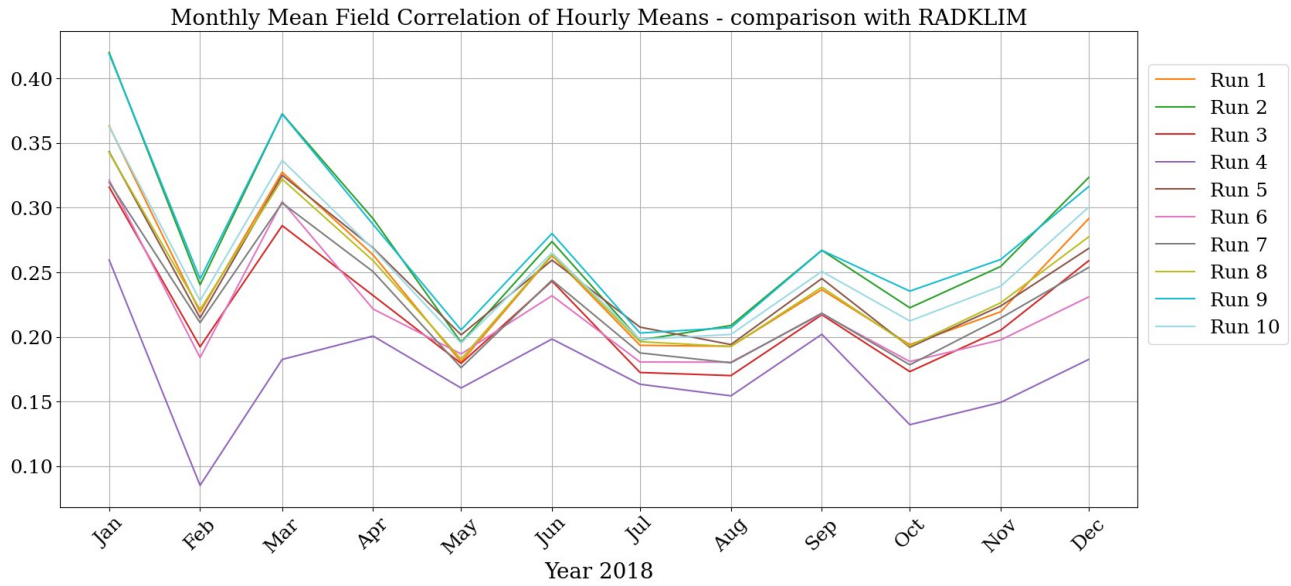


Figure 3.5: Monthly mean Field Correlation of the year 2018 calculated at each grid point using RADKLIM as reference. Colored lines represent the model runs.

In addition to evaluating the correlations of my models with HYRAS and RADKLIM, I have also computed the field correlation between these two reference datasets, as depicted in Figure 3.4, akin to the approach in Figure 3.1. The correlation between HYRAS and RADKLIM is exceptionally high, establishing once again a benchmark for all my models. A key observation is that runs 9 and 2, both employing the multiple timestep training methodology, achieve the highest field correlation scores. This outcome underscores the significant value of incorporating temporal data from multiple timesteps, not only in accurately predicting the quantity of precipitation but also in precisely mapping its spatial distribution. This trend is consistent in my models' comparisons with both HYRAS and RADKLIM for the year 2018. Run 10 ranks as the third-best, which, like runs 9 and 2, integrates temporal information alongside pressure data. However, the addition of pressure data does not appear to be as important as the temporal elements. Following these top performers, runs 1, 5, and 8 exhibit moderate performance levels, with correlations ranging between 0.2 and 0.35, as indicated in both Figures 3.4 and 3.5. Meanwhile, runs 3, 6, and 7, while scoring lower than that, do not fall far behind. Conversely, run 4 consistently shows significantly lower performance, a trend that aligns with its RMSE scores as well. An encouraging conclusion is that all runs maintain positive correlation scores. However, it's crucial to recognize that even the best-performing Runs, 2 and 9, still fall short of the HYRAS-RADKLIM benchmark by a margin of about 0.2 to 0.3. This gap highlights the room for improvement in achieving the high standard set by the reference dataset correlation. Throughout 2018, the correlation

values for all models stay within a narrow range of 0.1 to 0.4. This consistency suggests that all months presented the same level of difficulty in being predicted.

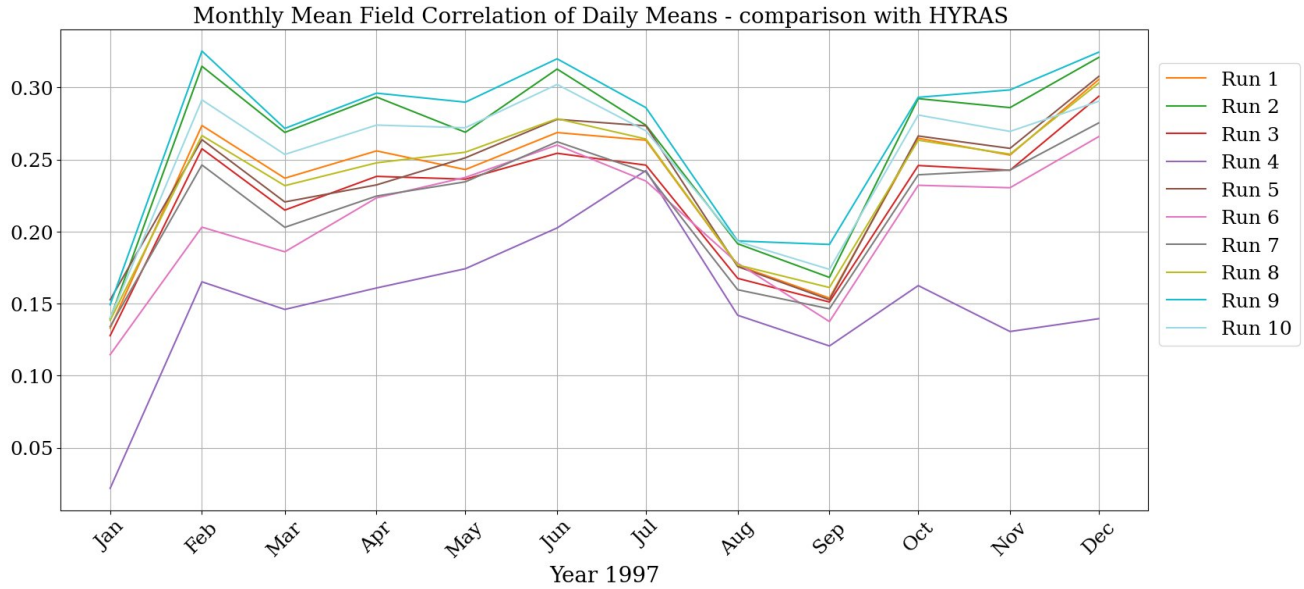


Figure 3.6: Monthly mean Field Correlation of the year 1997 calculated using HYRAS as reference. Colored lines represent the model runs.

In the case of 1997, a similar pattern to 2018 emerges, with runs 2 and 9 achieving the highest correlation scores, closely followed by run 10. The other models exhibit slightly lower performance, with their scores ranging between 0.11 and 0.3. Similar to the trend seen in 2018, Run 4 consistently shows the weakest correlation, occasionally dropping to values near zero, as observed in January. In contrast to the 2018 data, the predictive skill of the models in 1997 appears to fluctuate more significantly over the course of the year. Notably, all models experience a reduction in their correlation scores during January and September. However, for the remaining months, the models generally display higher correlation values, indicating varying levels of predictive accuracy across different times of the year.

Run configurations: 1 (precipitation, long train), 2 (precipitation, 2 timesteps, long train), 3 (precipitation, pressure change), 4 (precipitation, pressure, temperature, wind), 5 (precipitation, wind), 6 (precipitation, temperature), 7 (precipitation, pressure) ,8 (precipitation, short train), 9 (precipitation, 2 timesteps, short train) , 10 (precipitation, pressure, 2 timesteps) (see also Table 2.4)

3.3 Temporal Correlation Maps: Consistency of Model Predictions over Time

Continuing my investigation of the correlation between my models' reconstructions and the reference datasets, I now shift my focus from spatial to temporal correlation at each grid point. I use Equation 3.4 to calculate the Pearson correlation coefficient for each point across all timesteps. This approach allows us to assess the consistency of the relationship between two fields over time at specific locations. The result of this computation is illustrated in maps (Fig. 3.7 and Fig. 3.8), where each grid point depicts the temporal correlation between the datasets at that specific point. Unlike the monthly time series analysis discussed earlier, this temporal correlation investigation uncovers areas where a model may consistently overestimate or underestimate precipitation over time. By combining the insights from both field and time correlation coefficients, I gain a comprehensive understanding of model performance. This dual approach enables us to not only examine how accurately models capture spatial patterns at individual timesteps but also assess their ability to consistently replicate temporal patterns over an extended period, thus enhancing my spatial analysis with a crucial temporal dimension.

$$P_{time\ cor} = \frac{\sum_{i=0}^{W-1}\sum_{j=0}^{H-1} R_{t,i,j} O_{t,i,j} - N\,\overline{R_{i,j}}\,\overline{O_{i,j}}}{\sqrt{\left(\sum_{i=0}^{W-1}\sum_{j=0}^{H-1} R_{t,i,j}^2 - N\,\overline{R_{i,j}}\right)\left(\sum_{i=0}^{W-1}\sum_{j=0}^{H-1} O_{t,i,j}^2 - N\,\overline{O_{i,j}}\right)}} \qquad \text{(Eq. 3.4)}$$

Figure 3.7 depicts Time Correlation Coefficient Maps for runs 1-10, offering a detailed comparison across three columns. The first column displays the correlation between my models' output and HYRAS, and the second column represents the correlation with RADKLIM. The third column features a single map illustrating the correlation between HYRAS and RADKLIM, setting a benchmark for my models. This benchmark is characterized by high correlations, typically over 0.9, except for a linearly shaped region extending from central to northeast Germany with lower values. My models do not exhibit this weak correlation pattern, possibly due to the presence of multiple stations in that area. A distinct triangular area in Northeast Bavaria shows lower correlations, nearing zero, in maps comparing RADKLIM with other datasets. This pattern, resembling a radar's circular field, suggests a radar measurement issue absent in my model's output.

Focusing on the first column of Fig. 3.7, I observe high correlations across all models, indicating their proficiency in replicating the temporal variability of precipitation. However, a specific area in northwestern Lower Saxony presents challenges, with runs 1, 2, and 7 showing near-zero correlation,which means that the models there may not

underestimate or overestimate precipitation, but they are generally not reliable. In contrast, runs 5 and 10 display correlations close to -1, which shows that the model's predictions consistently and exactly oppose the observed temporal trend in the reference dataset. For precipitation, this means that when the model predicts high precipitation, the actual observation tends to be low, and vice versa. Runs 3 and 4, while not specifically weak in that area, display random zero-correlation regions, appearing as artifacts on the map. Runs 6, 8, and 9, however, yield highly correlated maps.

The second column of Fig. 3.7 reveals a white line in all maps across southwestern Baden-Württemberg. This is the result of a discrepancy present in the radar, and can be also detected in the RADKLIM data, in every timestep. This results in zero-correlation areas in my models, which do not detect this anomaly as precipitation. Additionally, border areas show problematic correlations, exhibiting either zero or negative correlation. This is to be expected, since the boundaries of the RADKLIM grid often fall outside of Germany, in areas where stations might be present, but were not used in this project. Generally though, the discrepancies are more evident here than they were on the correlations calculated between my models and HYRAS, often appearing as noise on the map, as in runs 3,4,8 and 10. Despite these discrepancies, high correlations are observed in most areas.

Run configurations: 1 (precipitation, long train), 2 (precipitation, 2 timesteps, long train), 3 (precipitation, pressure change), 4 (precipitation, pressure, temperature, wind), 5 (precipitation, wind), 6 (precipitation, temperature), 7 (precipitation, pressure) ,8 (precipitation, short train), 9 (precipitation, 2 timesteps, short train) , 10 (precipitation, pressure, 2 timesteps) (see also Table 2.4)

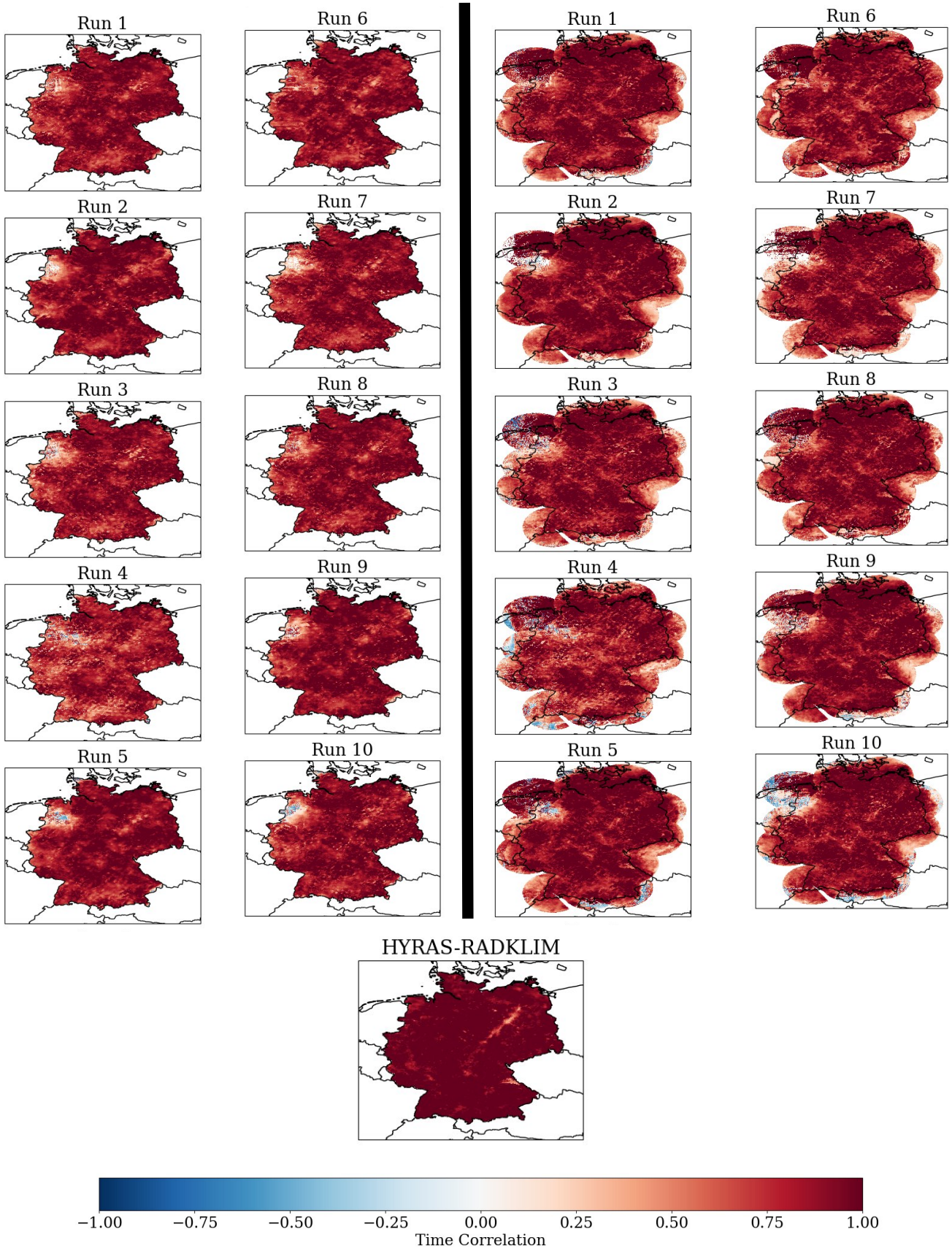


Figure 3.7: Time Correlation Maps for runs 1-10 depicting time correlation between the models' output and HYRAS (first and second column), the models' output and RADKLIM (third and fourth column), as well as the pair HYRAS-RADKLIM (bottom) for 2018

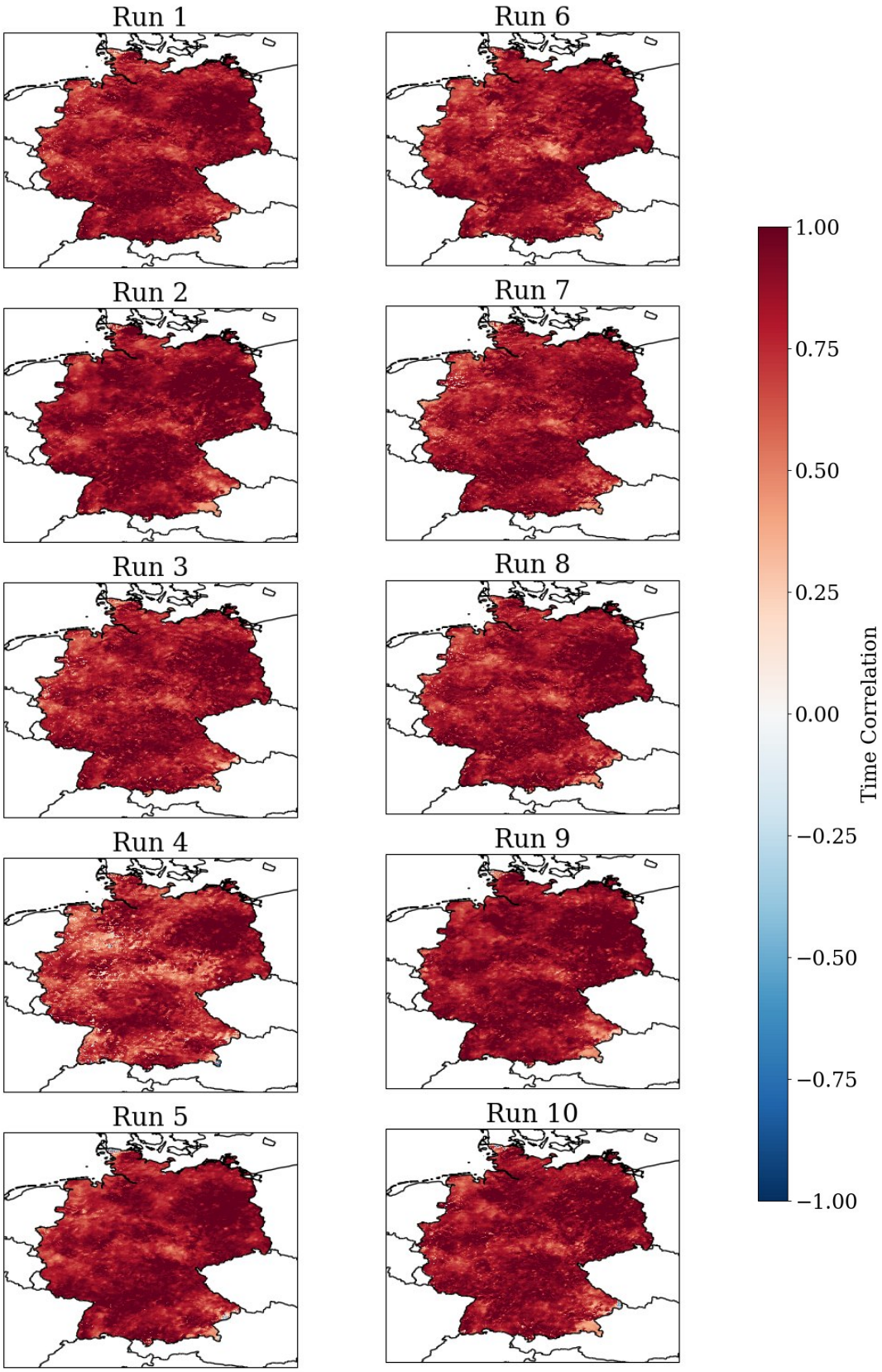


Figure 3.8: Time Correlation Maps for runs 1-10 depicting time correlations between my models and HYRAS for 1997

Run configurations: 1 (precipitation, long train), 2 (precipitation, 2 timesteps, long train), 3 (precipitation, pressure change), 4 (precipitation, pressure, temperature, wind), 5 (precipitation, wind), 6 (precipitation, temperature), 7 (precipitation, pressure) ,8 (precipitation, short train), 9 (precipitation, 2 timesteps, short train) , 10 (precipitation, pressure, 2 timesteps) (see also Table 2.4)

In Fig. 3.8, depicting 1997 correlations between my models and HYRAS, all maps show high correlations without negative values. Run 4 has several regions with correlation close to 0, that appear as checkerboard artifacts similar to these observed in 2018. Runs 2 and 9 excel in accuracy, with run 9 exhibiting fewer artifacts. While no clear advantage is discerned across the runs, certain regions show consistently better correlations: mid south Germany, east Germany, as well as some isolated locations near to the borders with Switzerland. In central Germany, there's a specific zone where all runs exhibit correlation values closer to 0. However, I find that this is a matter of extreme detail; overall, the performance of the runs is objectively good.

In summarizing the insights from the Time Correlation maps, a key observation is the absence of distinct patterns when comparing the 1997 and 2018 maps. I am unable to draw confident conclusions about the models' location-specific predictive strengths or weaknesses. Interestingly, I note an increase in accuracy in areas with denser station coverage, as seen in western Germany during 1997. Considering both temporal and spatial correlation analyses, it is evident that runs 2 and 9 stand out as the most effective. This outcome is expected, particularly for temporal correlation, given that these two runs utilize the multiple timesteps training methodology.

3.4 Difference of Total Precipitation between Model and Reference Data

In this section I investigate the final metric, the difference in total summed precipitation. This metric quantifies the overall difference in precipitation as recorded by two datasets. A positive value would suggest that the reference dataset shows more total precipitation than my model over the considered period, while a negative value would indicate that my model underestimates the precipitation amounts.

First I sum up the precipitation data spatially in each timestep for the model output and the two reference datasets. Next I calculate the sum over time for each precipitation field and thus I end up with the total precipitation amount over the entire spatial domain and throughout the entire time period covered by each dataset. I then subtract the total precipitation sums of the model's output from the reference dataset sums. These calculations are described by Eq. 3.5. The results, shown in Table 3.2, are single values representing the total difference in precipitation amounts between the references over the entire area and period covered. I also calculate the P_{diff} between the two reference datasets by subtracting the HYRAS field and the RADKLIM field.

$$PR_{diff}(R,O)=\sum_{t=0}^{N-1}\sum_{i=0}^{W-1}\sum_{j=0}^{H-1}R_{t,i,j}-\sum_{t=0}^{N-1}\sum_{i=0}^{W-1}\sum_{j=0}^{H-1}O_{t,i,j} \qquad \text{(Eq. 3.5)}$$

In line with previous sections, the score calculated between the two reference datasets, RADKLIM and HYRAS, yields the best results. Specifically, the total precipitation difference between RADKLIM and HYRAS is approximately $3 \cdot 10^5$ mm/h. When comparing my model outputs to these references, the precipitation differences generally reach magnitudes of 10^6 mm/h, suggesting a significant underestimation of precipitation by my models. Unsurprisingly, runs 9 and 2 emerge as top performers, reinforcing the conclusions drawn from other metrics previously discussed. Run 10 ranks third, closely following run 2, with all three of these runs employing multiple timesteps in their predictions. Following these, runs 1, 3, and 5 show adequqate performance. In line with other metrics, run 4 is identified as the least accurate, at times reaching error magnitudes of 10^7, especially in comparisons with RADKLIM.

A notable result is that the scores achieved through comparison with RADKLIM are almost twice the scores achieved in the HYRAS case for 2018. This can be attributed to the higher resolution of the original RADKLIM grid compared to HYRAS, allowing more room for errors and inaccuracies in the model's predictions. Additionally, it is interesting that the errors for 1997 are consistently higher than those for 2018 by an error rate of around $0.8 - 1 \cdot 10^6$ mm/h.

Run configurations: 1 (precipitation, long train), 2 (precipitation, 2 timesteps, long train), 3 (precipitation, pressure change), 4 (precipitation, pressure, temperature, wind), 5 (precipitation, wind), 6 (precipitation, temperature), 7 (precipitation, pressure) ,8 (precipitation, short train), 9 (precipitation, 2 timesteps, short train) , 10 (precipitation, pressure, 2 timesteps) (see also Table 2.4)

Run	$\mathbf{P_{diff,HYRAS,2018}} \cdot \mathbf{10^6}$ **(mm/h)**	$\mathbf{P_{diff,RADKLIM,2018}} \cdot \mathbf{10^6}$ **(mm/h)**	$\mathbf{P_{diff,HYRAS,1997}} \cdot \mathbf{10^6}$ **(mm/h)**
1	4.9371	9.8548	5.9799
2	**3.1897**	**7.2481**	**3.9035**
3	4.5321	9.2866	5.4600
4	6.7232	12.2705	8.1331
5	3.8204	8.3847	4.6019
6	5.1772	10.2371	6.2181
7	5.1200	10.1036	6.1945
8	4.9324	9.80848	5.9561
9	**2.8747**	**6.7933**	**3.3587**
10	3.1937	**7.3447**	**3.8451**
RADKLIM	**0.3648**	-	-

Table 3.2: Difference of Total Precipitation (P_{diff}) between my output and HYRAS for 2018 (first column), my output and RADKLIM for 2018 (second column) and between my output and HYRAS for 1997 (third column). The last row shows the quantities calculated between HYRAS and RADKLIM for 2018. The values in bold represent the best scores for each column.

4. Discussion

In this chapter, I determine which model is the most accurate based on the performance results from Chapter 3. The selected model is further discussed and then applied to reconstruct two periods; the years 1995-2000, which is the chronological gap in past precipitation recordings that my thesis aims to fill, as well as the years spanning 2008-2022. I further explore specific weather events from 2021, 2011, and 1999 to assess if my model reconstruction accurately describes these occurrences and improves upon existing records. The analysis focuses on the representation of these events in my model, compared to reference data. This comparison is crucial for evaluating the precision of my model in capturing the complex nature of precipitation, validating the physics of my reconstructions, and identifying potential strengths and weaknesses in my methodology.

4.1 Model Selection through Performance Evaluation

Among my various model runs, 9 and 2 consistently outperform others. This is an interesting finding since these runs employ the multiple timestep training methodology, but no additional meteorological variables, or quantities associated with them (like the pressure temporal change used in run 3). This suggests the significant impact that temporal information has on predicting precipitation. Run 10, which also uses temporal data along with pressure information, follows closely behind. Its relatively good performance reinforces the argument that incorporating temporal information is crucial in accurately capturing precipitation patterns.

The addition of pressure in run 10 does not show a marked improvement over 9 and

2. This observation leads us to question the value of different types of meteorological data in improving model accuracy. It seems that while additional data, like pressure or wind, provide some benefits (for example run 5, which uses wind, exhibited particularly good performance according to the RMSE scores), their positive impact is not as pronounced as the integration of time. This is observed also in the performance of runs 1 and 3, that display adequate performance but definitely not as commendable as run 2 and 9.

The most interesting result is the weak performance of run 4. This run, while it integrates a variety of station measurements, including wind,temperature, pressure and precipitation, does not perform as well as runs 2,9 or 10. This suggests that merely increasing the variety of data inputs does not guarantee better performance. It might be that the integration of multiple data types requires a more refined approach to model training and data processing. Additionally, the present modeling framework may have limitations in efficiently incorporating multiple data inputs. To overcome this obstacle, modifications in the code are nessecery.

In my analysis, notable discrepancies emerge when comparing the performance of my models against the RADKLIM and HYRAS datasets for 2018. Further, a comparison of metrics between 2018 and 1997 reveals distinct performance variations. Metrics such as RMSE, field correlations, and differences in total precipitation consistently indicate a relatively lower performance for 1997 compared to 2018, which could be realted to yearly variability.

It becomes evident that runs 9 and 2, by integrating multiple timesteps, stand out as the most effective models in my study. This effectiveness is not just limited to capturing the quantity of precipitation but extends to accurately predicting its spatial and temporal distribution. The superiority of these runs is consistent across different metrics and reference datasets. The two runs have similar performance with no clear advantage, proving that the performance is not related to the size of the training dataset. Among the two runs, I choose run 9 as the best, because by utilizing less data it is faster to train, making it more efficient than run 2 (run 9 needs almost 24 hours to reach 80000 iterations, compared to 96 hours that are needed for run 2). Moreover, since its training and validation dataset size is smaller, run9 is less expensive to use with respect to data storage.

It is important to acknowledge that my performance evaluation procedure is based on the assumption that HYRAS and RADKLIM accurately depict true precipitation fields. However, in reality, these datasets may not necessarily be the most physically accurate

Run configurations: 1 (precipitation, long train), 2 (precipitation, 2 timesteps, long train), 3 (precipitation, pressure change), 4 (precipitation, pressure, temperature, wind), 5 (precipitation, wind), 6 (precipitation, temperature), 7 (precipitation, pressure) ,8 (precipitation, short train), 9 (precipitation, 2 timesteps, short train) , 10 (precipitation, pressure, 2 timesteps) (see also Table 2.4)

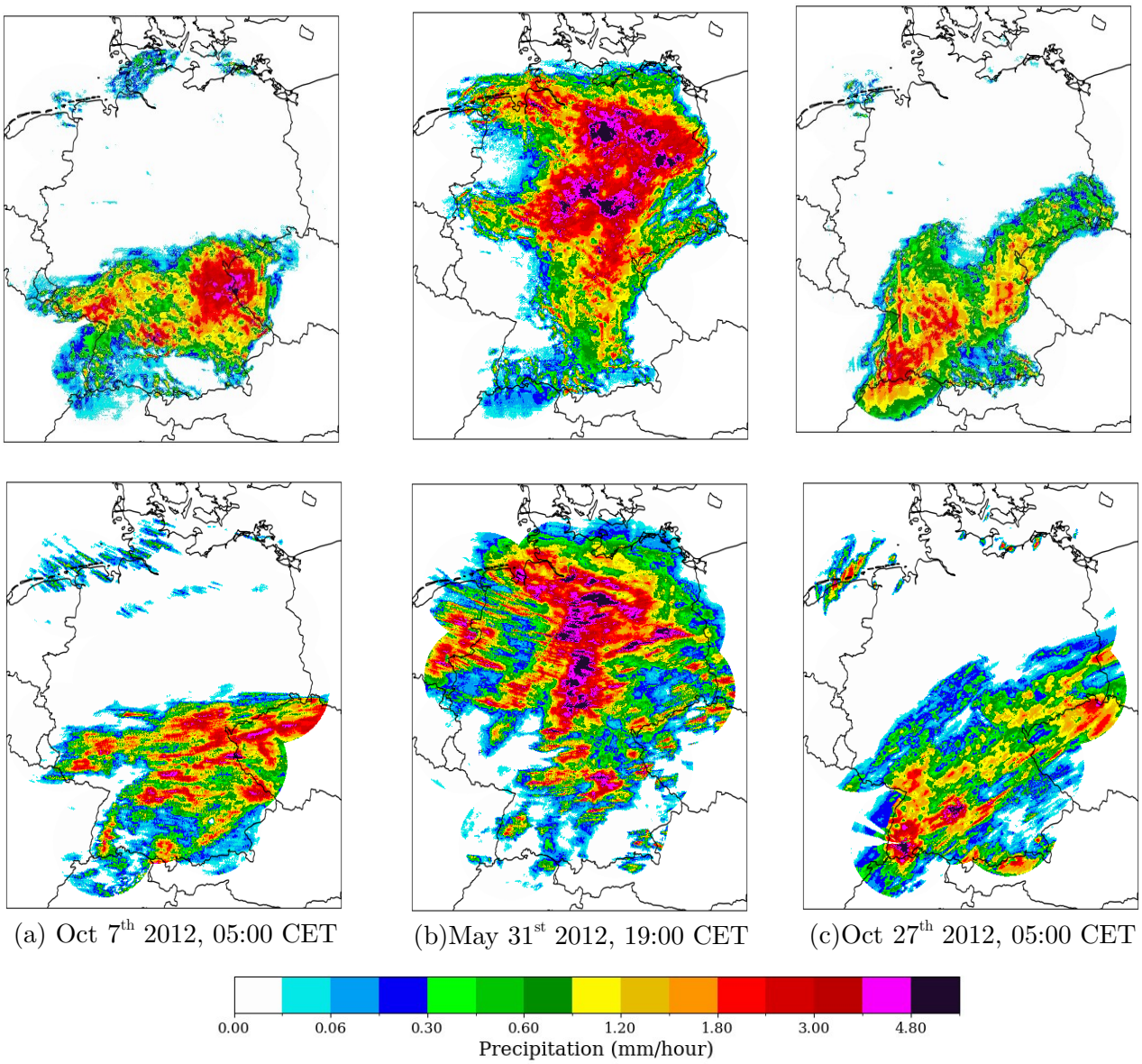
representations of actual precipitation and they each have their limitations. Nevertheless, they serve as my best available benchmarks for evaluating radar-based precipitation models, especially considering the absence of a precise spatial precipitation field for years prior to 2000. The fact that I rely heavily on these reference datasets as ground truth data introduces inherent limitations to my methodology, potentially impacting the physical interpretation of my conclusions. It is beyond the scope of this thesis to address these limitations. However, this highlights the need for continuous refinement and validation of precipitation models against various observational datasets to ensure their robustness and accuracy.

4.2 Application of Trained Model: Precipitation Reconstruction

In this section I use my best-performing model, as identified in Section 4.1, to reconstruct precipitation. The model receives as input the historical stations measurements and recreates a complete precipitation field. A large amount of data was created and therefore, only representative examples can be shown and discussed here. In Section 4.2.1 I present precipitation fields which correspond to hourly timesteps from the years 1997 and 2012. These maps are crucial for understanding the distribution and intensity of rainfall across different regions. In section 4.2.2 I show the monthly averaged time series for the reconstructed precipitation fields of September 1995 – December 2000 and January 2008 – December 2022. In these plots I can examine whether the model is able to capture temporal variations consistently over an extended period of a few years.

4.2.1 Investigation of 2012 and 1997 Precipitation Fields

Here, I investigate rain events over the course of a few hours. First, I reconstruct some timesteps from 2012 and I compare them to the corresponding hourly RADKLIM grids (Figure 4.1). I choose timesteps from October and May 2012, in the early morning and afternoon hours.



(a) Oct 7th 2012, 05:00 CET (b)May 31st 2012, 19:00 CET (c)Oct 27th 2012, 05:00 CET

Figure 4.1: Plots of precipitation fields for selected timesteps in 2012. The top row (a) Oct 7, 05:00, (b) May 31, 19:00, and (c) Oct 27, 05:00, depict the reconstructed precipitation fields. The bottom row presents the corresponding ground truth data taken from the RADKLIM dataset.

The model is capable of accurately representing the overall pattern of precipitation, particularly when it is concentrated in a specific region, as demonstrated in the examples of the first and third column. The model captures variations within that pattern, as indicated by the coherent structures of precipitation intensity, resulting in a physically plausible field that closely resembles the ground truth RADKLIM data. Furthermore, the model accurately reproduces the orientation and intensity gradient of the precipitation fronts. Outside of the German borders my fields' surface is limited, since I did not use stations that are located outside of Germany.

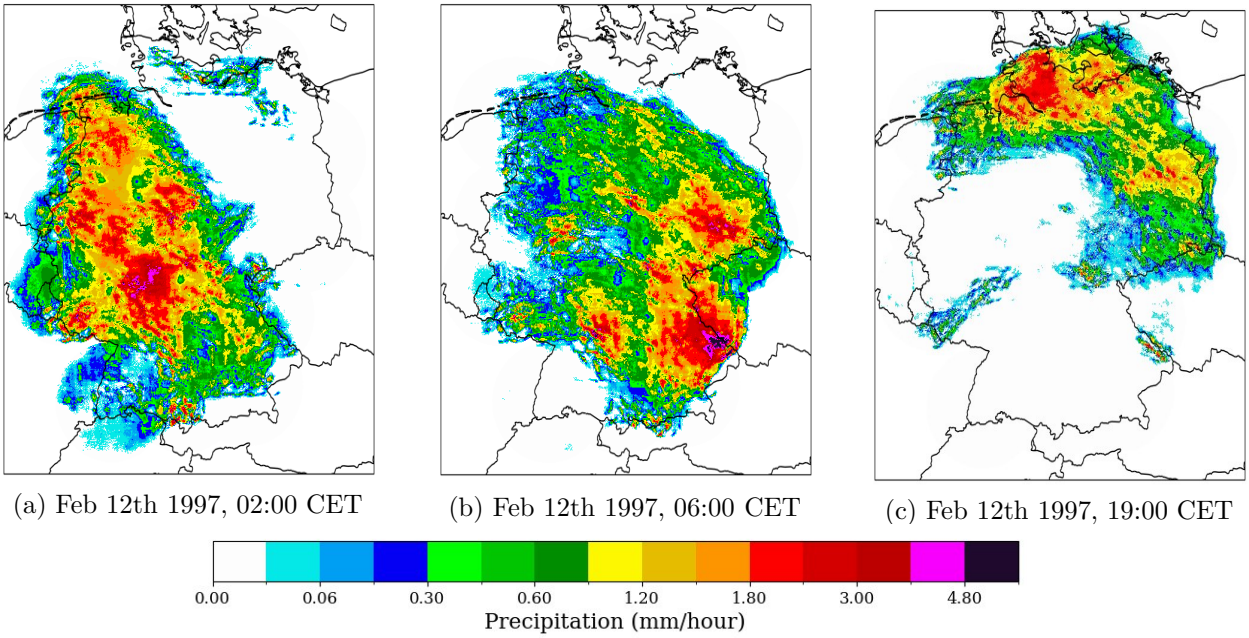
The model performs better in regions with denser station coverage, such as North-east Germany, where it can accurately replicate patterns of heavier precipitation ($>$ 3.2 mm), as shown in the plots in the second column. Despite this skill, there are issues with the misplacement of precipitation events in several locations, particularly evident in the second example where several heavy events occurred. The locations where the model predicts heavy rain are slightly different from where the radar actually recorded it. This can be attributed to the fact that the station coordinates (and thus the information that is available to the model to start reconstructing from) have a subtle deviation from the corresponding locations of the radar field measurements. When placing a station on the RADKLIM grid, if the exact coordinates were not available (due to resolution differences), the nearest available coordinates within a 1-kilometer radius were used. This approach, while practical, introduces a spatial discrepancy that could explain why the model misplaces certain precipitation events.

My outputs further exhibit several checkerboard pattens that appear on the edges of the precipitation fields. I have already encountered these peculiar shapes in the metric maps. They also make an appearance in the outputs of the original publication by Liu et al. (2018) and the radar infilling by Meuer et al. (2022) but they are not reported in the temperature grid reconstruction by Kadow et al. (2020). It is unclear where these artifacts stem from. The fact that they are so evident in all the works apart from Kadow et al. (2020) could be related to the higher resolution of the images. Previous works (Johnson et al. 2016) suggest that perceptual loss is often responsible for generating these checkerboard artifacts in real-world images, and the effect can be countered by using the total variation loss. In this study I use both losses but still encounter artifacts, similarly to Liu et al. (2018).

Next, I examine the timesteps in more detail. In the first timestep, there is a broad band of precipitation across the central region. Both the model output and RADKLIM indicate higher intensity areas (>2.8 mm) within this band, but the model appears to show a slightly broader area of moderate precipitation (between 0.8 and 2.8 mm), particularly in the eastern part. The model accurately reproduces the intensity of a significant event, which is the precise reconstruction of a small area with the highest intensity ($>$ 4.8 mm) on the border with the Czech Republic. The only difference is that it spreads the event over a larger area. The model overestimates while RADKLIM underestimates precipitation. This is because radars cover a larger area (1 km^2 per pixel) and record an average of the precipitation over that entire area. Although the radar has a higher likelihood of detecting precipitation within this larger coverage area compared to a point-based gauge, the spatial averaging can result in lower recorded intensities. This limitation is particularly relevant for heavy rainfall events, where localised intensities may exceed the average quantity across a radar pixel. Referring back to the first plot, some of the areas with the highest intensity (>4 mm) do not match perfectly; the model appears to misplace some events. Conversely, in the northern part of Germany, my model reproduces a small amount of rain, consistent

with the radar data. In the second example, both the model output and RADKLIM show a clear and intense storm system over the central region of Germany. As mentioned above, the model significantly misplaces the most intense events, except for one event that is accurately placed on the western side. In both the first and second examples, the ground truth radar data shows radar noise appearing as lines of precipitation, which does not correspond to physically acceptable precipitation patterns. My models do not reproduce these patterns. In other examples (not shown) similar unwanted noise that comes in the form of odd triangular forms or still pixels, is also not present in my model's output. In the final example, a small amount of precipitation is observed far from the northwest coast, which the model fails to capture due to the absence of stations in that area. The reconstruction is problematic in several locations, as the radar shows higher precipitation events that are not reflected in the model's output. The model does not capture the higher precipitation events detected by the radar, instead, it depicts a more distinct region of moderate intensity (between 1.6 and 2.8 mm). Moreover, it incorrectly recreates an area of around 2.4 mm in the border of the Czech Republic and overlooks another area of moderate precipitation on the western side.

Next, I reconstruct timesteps from 1997 and use HYRAS as a benchmark, since this is the selected reference dataset available for that time period. As a first step, I visualize hourly timesteps, selected from a single day that has been reconstructed by my model. This investigation allows us to closely evaluate the model's performance on an hourly basis. Next, I compute the daily average precipitation from my model's output. This daily average serves as a basis for comparison against the daily data from HYRAS.

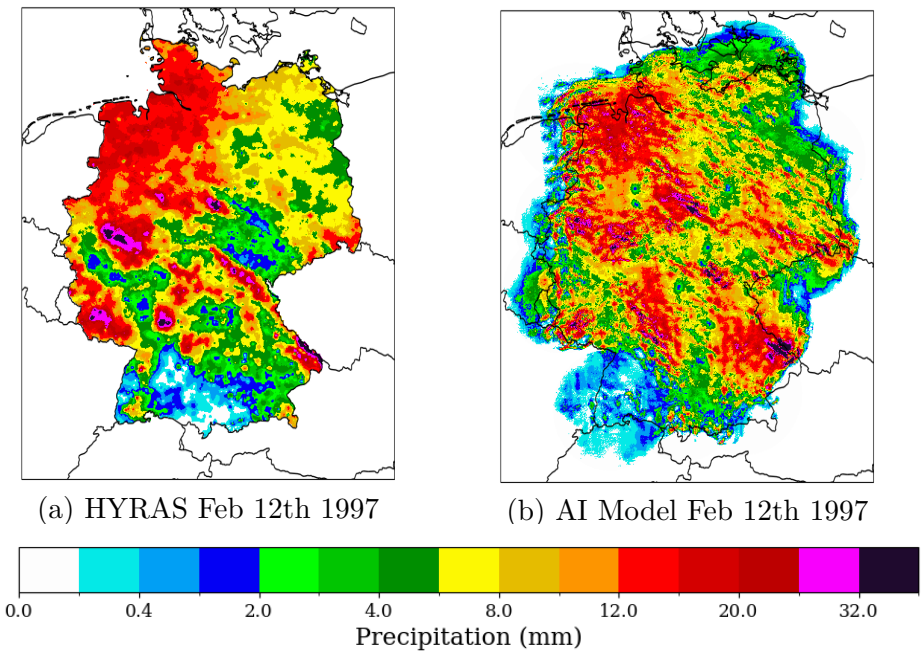


(a) Feb 12th 1997, 02:00 CET (b) Feb 12th 1997, 06:00 CET (c) Feb 12th 1997, 19:00 CET

Figure 4.2: Plots of reconstructed precipitation fields for selected hourly timesteps on 12/2/1997: (a) 02:00, (b) 06:00, and (c) 19:00

The first day of study is the 12th of February (Fig. 4.2). In all three plots that correspond to different times of the day the checkerboard artifacts are evident once again, mainly on the edges of the precipitation patterns. In the first plot I see a significant precipitation event centered over the middle of the country. Small 'hotspots' of high precipitation ($>$ 2.8 mm) can be detected. Certain high precipitation locations align with the positions of stations particularly in central Germany. Remarkably, in the model output I can also detect similar high-intensity events that have a very narrow spatial spread, but are not close to any stations. For instance, a localized event within central Germany that resembles a single data point is captured by the model, showing precipitation levels surpassing 4 mm. Moving from the center towards the south, a similar event can be observed in the second timestep. Moreover, in the second timestep I see an intense pattern ($>$ 4 mm) near the border with the Czech Republic that can also be detected both in the HYRAS data and in my daily averaged output (Fig. 4.3). In the third timestep I see a uniform precipitation field of moderate intensity in the north. If I further compare the daily average plots I conclude that even though my output has a lot of checkerboard effects (aggregated from the hourly timesteps) it is still able to resemble HYRAS in several high intensity events. In the south, my output misplaces a small amount of precipitation, moving it towards the east. Slight differences in the location of events are also evident in the rest of the map. A possible explanation for this stems from the grid difference, similarly to the 2012 comparison; my output employs the RADKLIM grid, which is defined differently than HYRAS, even though they have the same spatial resolution. Nonetheless,

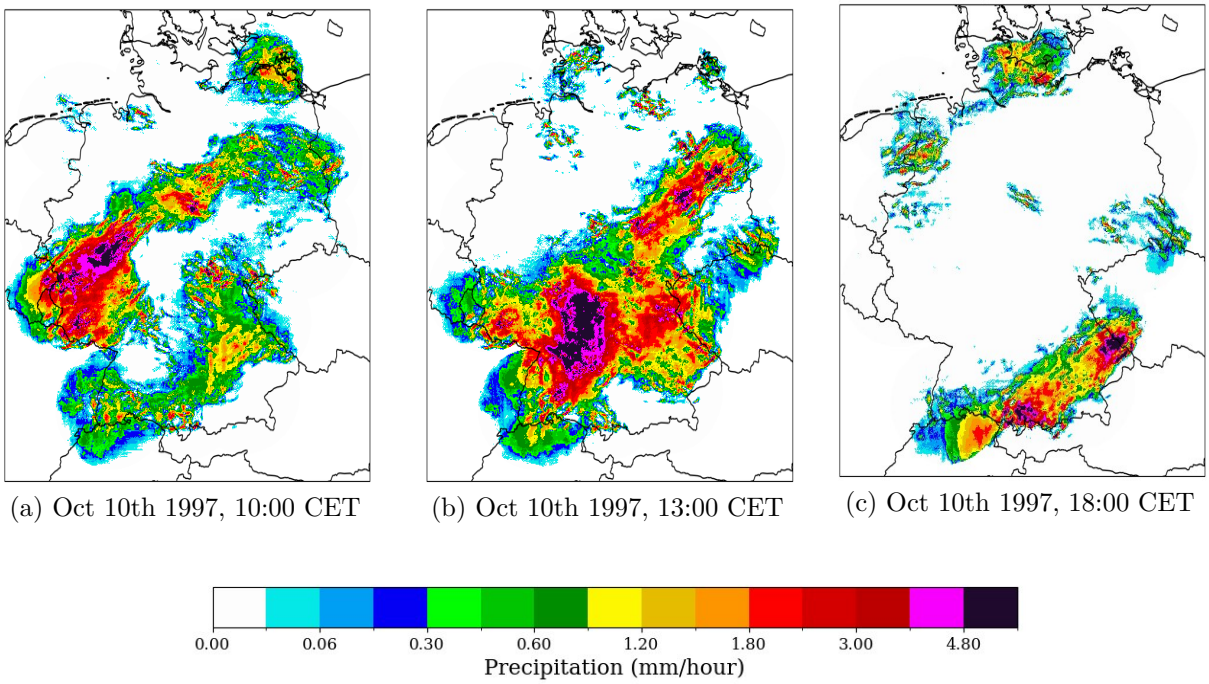
both plots capture the day's significant rainfall event, but the model appears to have a higher variability in precipitation intensity over small distances.



(a) HYRAS Feb 12th 1997 (b) AI Model Feb 12th 1997

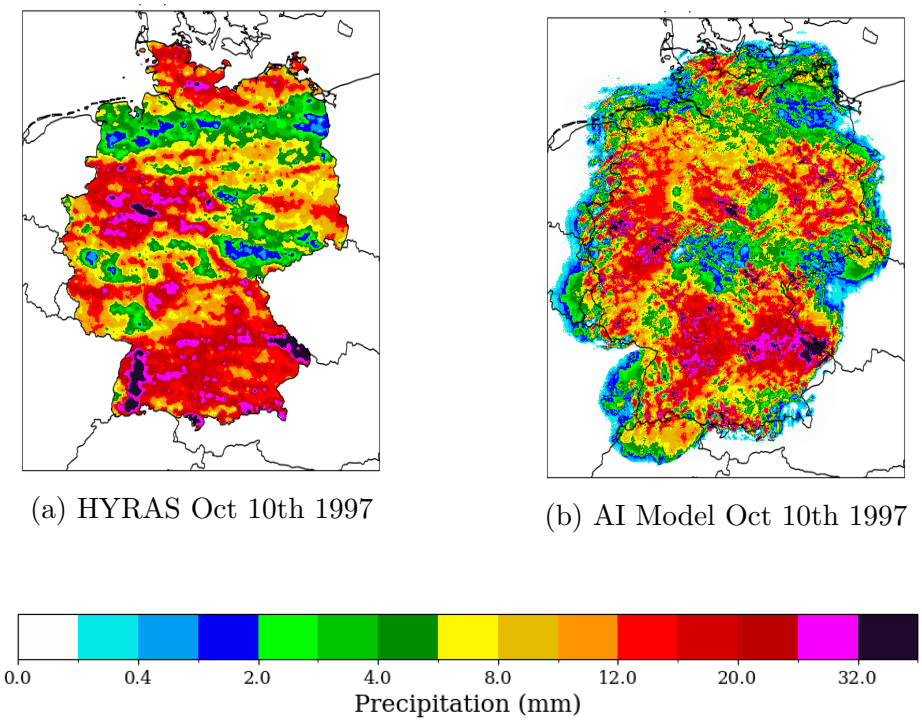
Figure 4.3: Precipitation daily averages of 12/2/1997 for (a) HYRAS-PRE (b) the AI model

The second case study from 1997 is the 10th of October (Fig. 4.4). In the first timestep I see a small-scale event located in the center of Germany, in the exact same location as a similar event on the first timestep of Fig. 4.5. A few kilometers southeast from that location, an extended checkerboard artifact is placed, that exhibits remarkably high intensity. On the west, a uniform pattern of high intensity is detected, that is intense enough to be found in the daily average plots (Fig. 4.5). However, in my model's daily averaged plot, its spread is very limited, compared to HYRAS. In the northeast, an unrealistic circular and patchy pattern is seen. Similar shapes are particularly evident on the third timestep of the same figure. The second timestep is governed by a strong event starting from the center of Germany and moving towards the southwest.



(a) Oct 10th 1997, 10:00 CET (b) Oct 10th 1997, 13:00 CET (c) Oct 10th 1997, 18:00 CET

Figure 4.4: Plots of reconstructed precipitation fields for selected hourly timesteps on 10/10/1997: (a) 10:00, (b) 13:00, and (c) 18:00



(a) HYRAS Oct 10th 1997 (b) AI Model Oct 10th 1997

Figure 4.5: Precipitation daily averages of 10/10/1997 for (a) HYRAS-PRE (b) the AI model

I encounter several smaller scale events of high intensity as well, that are averaged out and thus do not make an appearance the day average plots. The third timestep does not record a widely distributed amount of precipitation, with the exception of the southern part. There, a very intense event is detected. That event, along with others of equal intensity, are noted in both plots of Fig. 4.5, and they are placed by the model in locations that coincide with HYRAS.

Upon examining the daily averages for both days, I conclude that on the 10^{th} of October the model was not as skillful on reproducing the precipitation patterns, as it was on the 12^{th} of February. In the October timestep, the extremely intense event on the southwest is not detected. Moreover, in the north, the model underestimates the spread of an event that was particularly intense. Regardless, the fact that the model is able to recreate patterns that resemble HYRAS, even at its daily resolution, is a promising result. It suggests its capability to replicate both widespread as well as isolated precipitation patterns that are physically acceptable, despite a few drawbacks.

4.2.2 Time Series Investigation of Monthly Mean Precipitation

In this section, I calculate the monthly mean precipitation from my reconstructed precipitation output, and I compare them against the reference datasets HYRAS and RADKLIM. Reconstructions start only from September 1995 and not earlier because hourly station measurements are available to us from that moment on. I exclude the years 2001-2007 from my reconstructions because they were part of the training and validation datasets.

Figure 4.6 illustrates the comparison between the monthly mean precipitation of my reconstructed output and the values derived from HYRAS. There is a notable agreement in the overall trend and the seasonal fluctuations, with both datasets demonstrating similar patterns of peaks and troughs throughout the 1995-2000 period. Despite this coherence, there is a consistent underestimation by the model, with a bias varying from 0.01 to 0.07 mm across different months. The model exhibits a stronger alignment with HYRAS during periods of lower precipitation, effectively capturing values below 0.03 mm, as seen in September 1995, January 1996, January 1997, and February 1998. However, the model's limitations become apparent during high-intensity events, where the most pronounced discrepancies are observed. This is particularly evident during months like June, July, and December 1997, October 1998, December 1999, and March and June 2000, where the model fails to accurately replicate the peak precipitation values found in HYRAS. One possible explanation for these successes and failures of the model lies in the training dataset; high-intensity extreme events are rare by definition, and as a result the model does not encounter them often during training. Low intensity events or events with no

precipitation at all are significantly more frequent. This imbalance means that the model's skill in detecting and quantifying extreme events is limited, explaining its tendency to underestimate them.

To delve deeper into the dry bias observed in the model, I conduct a value distribution analysis focusing on the precipitation intensity histogram for the year 1998 (Figure 4.7), where I plot the value distribution of the model's output and HYRAS. Consistent with the imbalanced training set, the histogram reveals a substantial prevalence of values near zero in the model's predictions, surpassing the corresponding counts in HYRAS. This prevalence of low-intensity to zero precipitation events is strongly associated with the dry bias seen in Fig. 4.6, and it is not limited to the year 1998. Specifically, the model tends to overestimate occurrences of such events compared to HYRAS. Conversely, when examining events with intensity higher than 0.1mm, the model consistently registers fewer counts than observed in the HYRAS data.

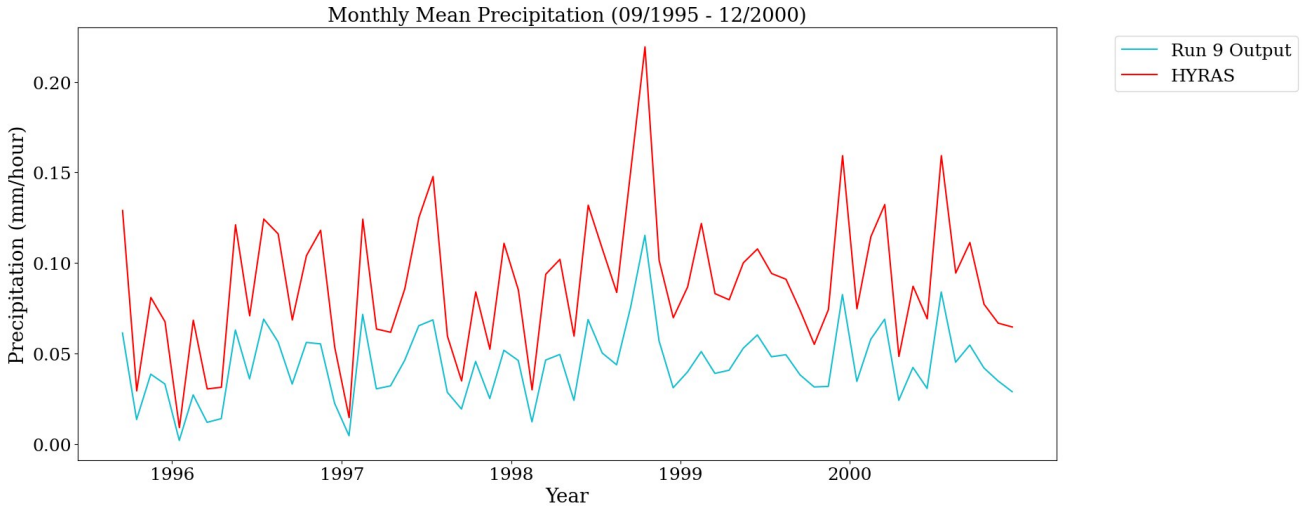


Figure 4.6: Monthly mean time series of the reconstructed years (light blue) and HYRAS (red) for the years 1995-2000, averaged over Germany

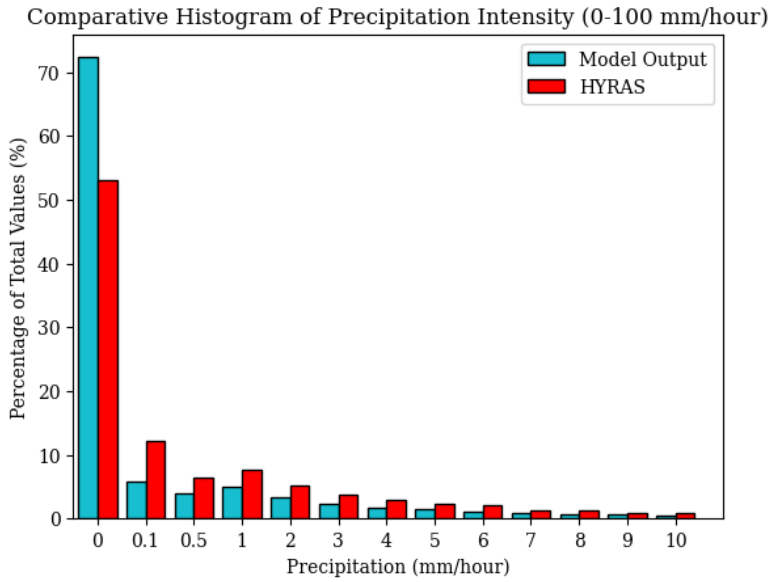


Figure 4.7: Histogram of precipitation intensity for the year 1998 for HYRAS and the AI model output. For the AI, more than 70% of the values are between 0 and 0.1 mm/hour, showcasing the imbalance in the dataset. In HYRAS the values are more evenly distributed.

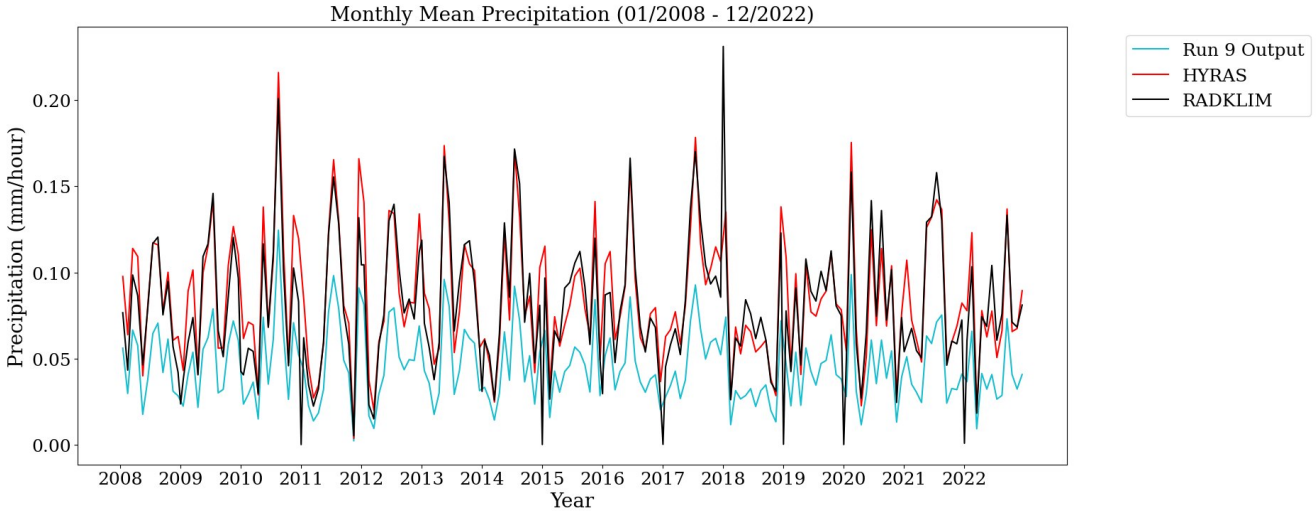


Figure 4.8: Monthly mean time series of the reconstructed years (light blue), HYRAS (red) and RADKLIM (black) for the years 2008-2022

The time series plot in Fig. 4.8 displays the monthly mean precipitation from 01/2008 to 12/2022. The plot includes data from the model output, HYRAS, and RADKLIM. Similarly to Fig. 4.6 the model output closely aligns with the peaks and troughs depicted by both HYRAS and RADKLIM, indicating a successful capture of the overall seasonal trends. Particularly the troughs are more accurately depicted by the model

than the peaks. There is again a notable bias in the magnitude of precipitation, particularly in the model's tendency to underestimate the higher-intensity rainfall events when compared to the reference datasets. The RADKLIM line occasionally diverges from HYRAS and the model output, which may be attributed to differences in measurement techniques— RADKLIM being radar-based, hence providing a broader spatial average of precipitation over its total area, as discussed before when investigating the precipitation maps. Throughout the span of the dataset, certain years stand out with pronounced peaks of precipitation, captured differently by the three datasets and suggesting differences in how each system records and interprets high-intensity rainfall.

The two plots of the monthly means offer some perspective on the performance of my model against HYRAS and RADKLIM. Although the comparison confirms a general agreement in capturing the broad trends of precipitation, the plots do not provide deeper insights on the model's output. The inherent averaging in monthly means masks the finer details and variations that are critical in understanding precipitation behaviors and oversimplifies the complexity of precipitation dynamics. The plots however validate that my outputs are physically plausible. In the next sections, I will follow an event-based approach and investigate specific events found in my model output, from a stricter meteorological perspective. In this way I hope to gain understanding on the representation of precipitation dynamics in my reconstructed fields.

4.3 Event-based Analysis of Past Precipitation Events

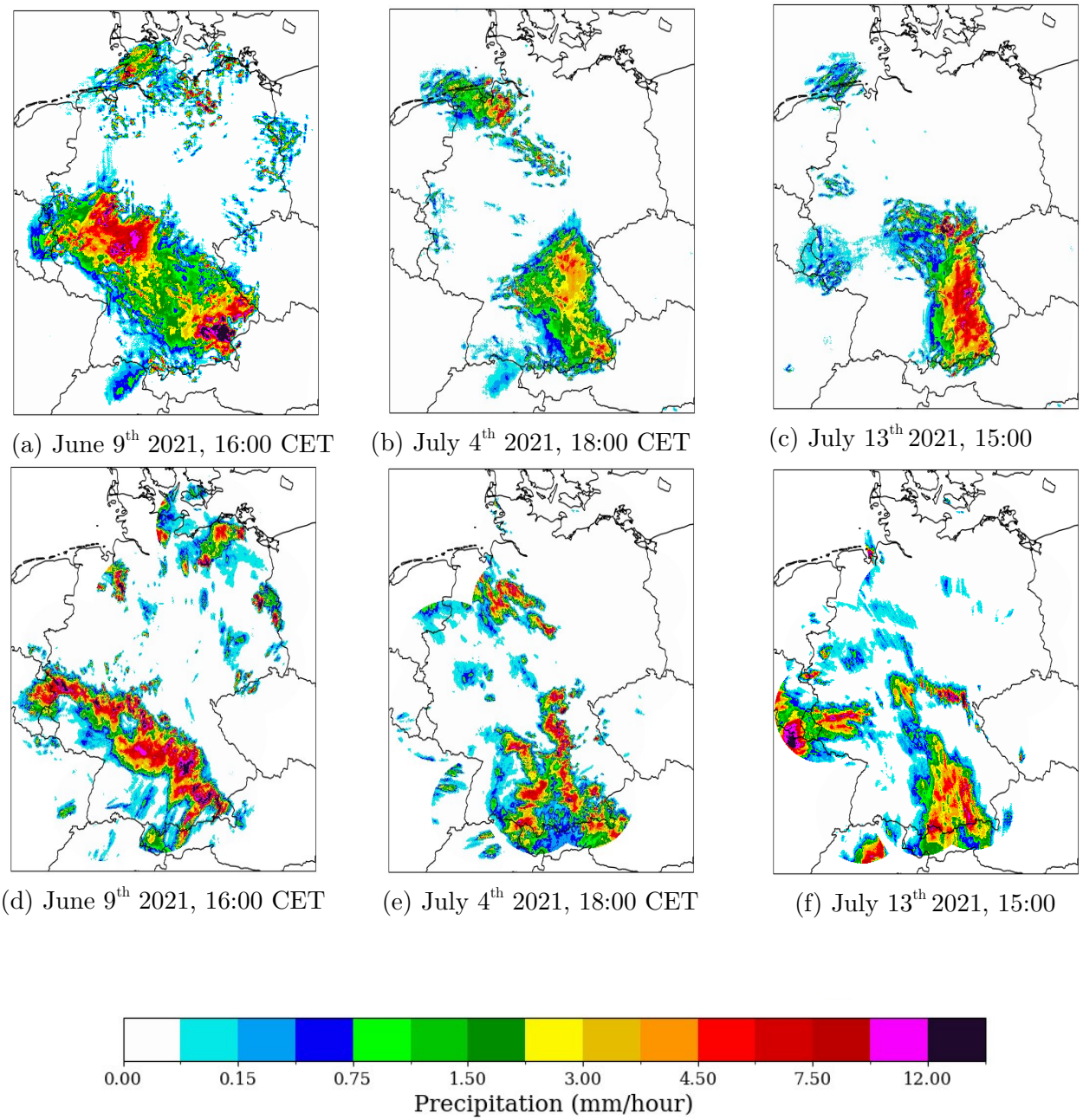
4.3.1 The 2021 Summer Floods in Germany

As I briefly mentioned in the introduction, the summer of 2021 in Europe was characterized by a series of intense precipitation events, largely influenced by the 'Bernd' storm system. This storm was a significant contributing factor to the catastrophic floods that severely impacted Germany and its neighboring countries. Given the extensive documentation and analysis of this event, it serves as an excellent reference to compare my model's output against. My goal here is not to surpass the accuracy of the data already used to study the event but rather use it as a benchmark to test how trustworthy my reconstructions are.

The year 2021 is among the top five years with the highest number of distinct intense precipitation events since 2001. In the three weeks leading up to the flood events, Germany experienced several heavy precipitation events, marking the wettest summer in

10 years and significantly saturating the soils all around the country. Consequently, according to Junghänel et al. (2021), in eastern Saxony, southeastern Bavaria and in the southwest of North Rhine-Westphalia the ground was partly saturated and it was able to absorb only limited amounts of water. In other regions of Bavaria, in almost all of Rhineland-Palatinate and Baden-Württemberg the soil was completely saturated. This effect combined with continuous intense precipitation resulted in significant surface runoff, which in turn led to floods.

The intense rainfall was part of a series of events that was ongoing since May 2021. Figure 4.9 shows two examples of such events that occurred in June 29th, July 4th and July 9th. In the first example, the model's reconstruction displays a broad swath of precipitation across Central and South Germany. The most intense precipitation with values reaching up to 15 mm in one hour is concentrated in the southwestern part of the country. Scattered areas of moderate precipitation are also present across the rest of the region. In the corresponding RADKLIM timestep in the bottom row a similar pattern is observed with a focus on the same geographical area. However, there are noticeable differences in the distribution and intensity of precipitation. The areas of highest intensity (with more than 12 mm) seem slightly less extensive and slightly shifted geographically compared to the model output in the first image. The second example in July shows a similar pattern. Both the model and RADKLIM exhibit a high concentration of precipitation in the central area but the spatial distribution is limited in the reconstructed field. In both examples however there is reasonable agreement between the precipitation fields. Notably, these events persist for hours or even for several days (not shown). When comparing the geographic distribution of rainfall indicated in these images with soil moisture data, I find a strong agreement with regions identified as having high soil saturation levels, as detailed in Junghänel et al. (2021). This suggests that these precipitation events, captured both by RADKLIM and my model, played a significant role in contributing to the soil saturation in those areas. A noted challenge with RADKLIM is radar malfunctions, leading to areas with no measurements on the map, as seen in Figure 4.9e) and f) where the Borkum radar malfunctioned. My model addresses this issue by infilling these regions, providing valuable insights into the event's manifestation in those areas.



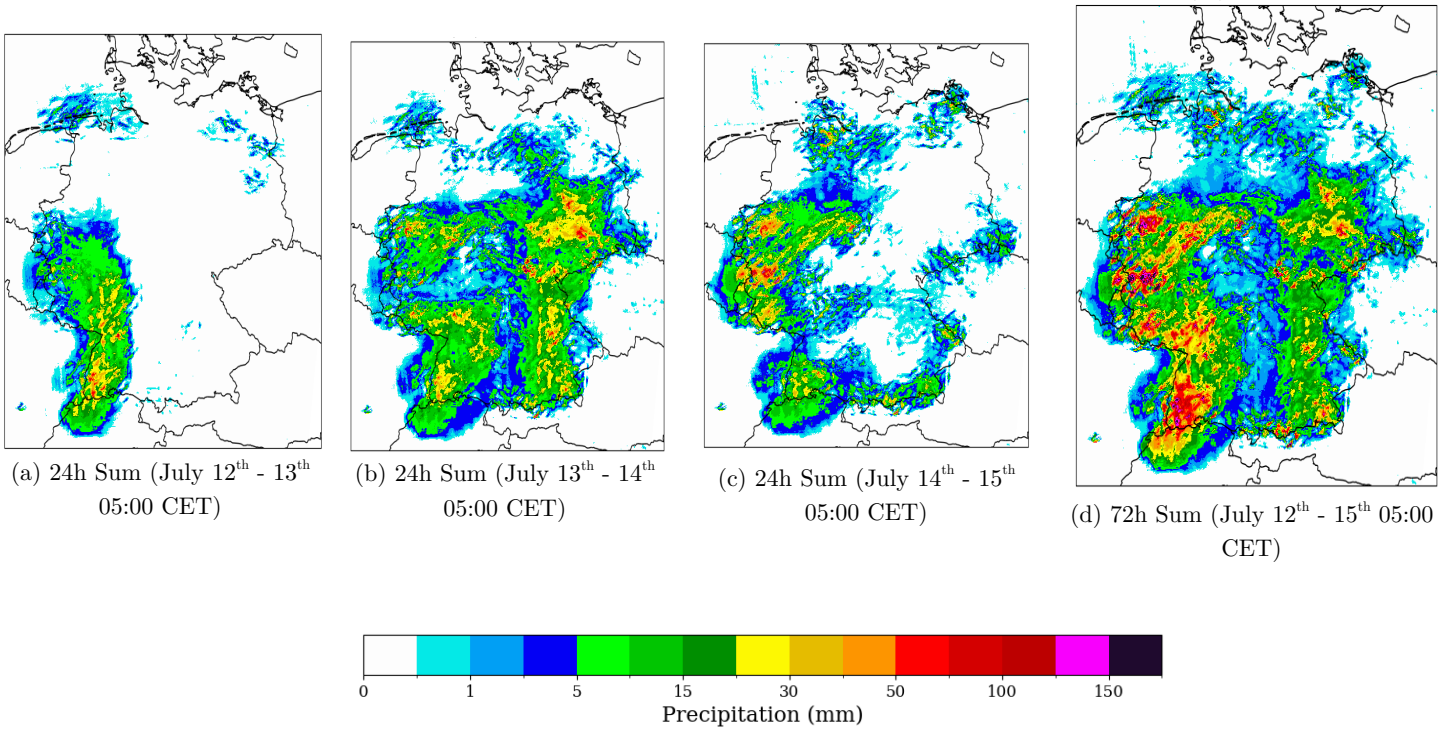
(a) June 9th 2021, 16:00 CET (b) July 4th 2021, 18:00 CET (c) July 13th 2021, 15:00

(d) June 9th 2021, 16:00 CET (e) July 4th 2021, 18:00 CET (f) July 13th 2021, 15:00

Figure 4.9: Plots of precipitation fields for selected timesteps in 2021. The top row displays reconstructed precipitation fields at specific times: (a) June 29th at 16:00, (b) July 4th at 18:00, and (c) July 13th at 15:00. The bottom row features the corresponding timesteps from the RADKLIM dataset.

To further investigate the event's development as described in the AI model's output, I calculate the 24-hour and 72-hour precipitation sums for 12, 13 and 14 of July 2021 (illustrated in Figure 4.10). This analysis enables a direct comparison with Fig. 1.1 and the contents of the corresponding DWD report. My model's depiction for the interval

of July 12-13 (Figure 4.10a) identifies discrete zones within Baden-Württemberg where precipitation exceeded 50 mm in a day. Extensive areas across Rhineland-Palatinate, Saarland, parts of Hesse, and North Rhine-Westphalia recorded considerable precipitation, predominantly surpassing 5 mm. These readings are part of a precipitation pattern that was initially concentrated in southwestern Germany, and over the course of a few hours, it moved towards central Germany. The subsequent plot (for July 13-14, Figure 4.10b) reveals this expansion of precipitation across Germany, with Baden-Württemberg still experiencing high levels. In most areas, precipitation reached 5 - 25 mm over 24 hours. Particularly high amounts (more than 75 mm) are observed around small regions in North Rhine-Westphalia, southern Bradenburg, Thuringia (close to the Selbitz station), central Bavaria, Saxony and along the German-Czech Republic border in the Ore Mountain range. As the storm abates on the third day, the model's snapshot for July 13 at 15:00 (Figure 4.9c) captures intense precipitation in southeastern Germany, significantly contributing to the daily totals presented in the subsequent figure. Isolated instances of substantial rainfall on July 14 (Figure 4.10c) are evident in the west, around Dortmund and Cologne.



(a) 24h Sum (July 12th - 13th 05:00 CET)

(b) 24h Sum (July 13th - 14th 05:00 CET)

(c) 24h Sum (July 14th - 15th 05:00 CET)

(d) 72h Sum (July 12th - 15th 05:00 CET)

Figure 4.10: Plots of 2021 precipitation sums for 24 hours (a) July 12th – July 13th at 05:00 CET, (b)July 13th – July 14th at 05:00 CET , (c) July 14th – July 15th at 05:00 CET and 72 hours (d) July 12th – July 15th at 05:00 CET. (similar to Fig. 1.1)

Overall the reconstructions agree with the RADKLIM ground truth data, capturing the long-scale patterns and trends of the weather event. A closer examination reveals that

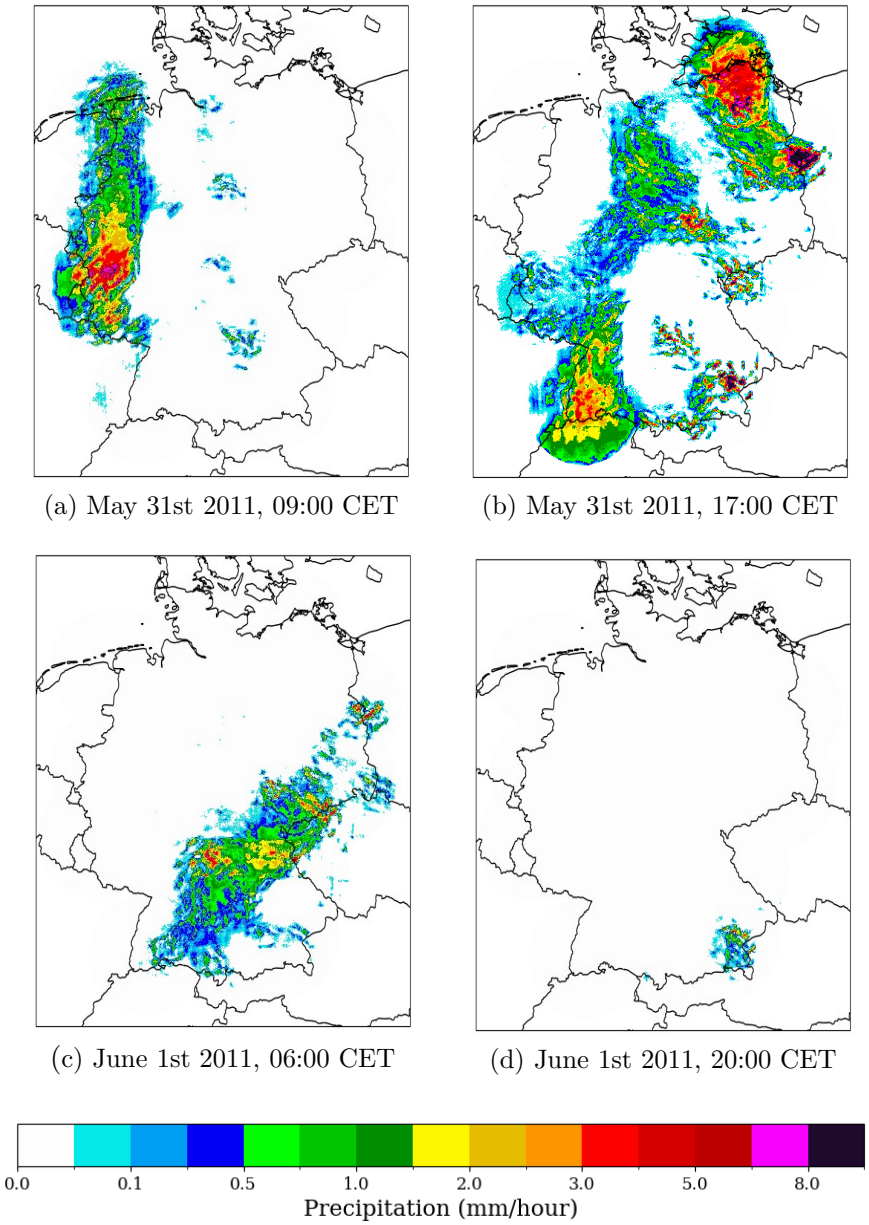
the checkerboard artifacts are more pronounced in the aggregated precipitation totals. Additionally, there is a noticeable spatial under-representation of rainfall, particularly concerning the breadth of the pattern. However, the magnitude of precipitation is accurately estimated, aligning closely with actual measurements. Although there are discrepancies in the precise localization of rainfall, key areas of intense precipitation are successfully identified. The model also adeptly traces the general trajectory of the storm. This severe event served as a valuable opportunity to assess the model's capability in simulating extreme weather phenomena.

4.3.2 Comparative Analysis of an Intense 2011 Event Using Reanalysis Data

At the end of May 2011 a cold front of significant size passed through Central Europe causing a rapid transition from very warm conditions to significant precipitation, leading to diverse meteorological impacts across different regions of Germany. This transition was particularly pronounced in southern and eastern Germany. For example, in northern Bavaria, some areas received over 70 mm of rainfall within two days. As mentioned in the official Early Warning Report published by the Karlsruhe Institute of Technology, the key feature of this weather pattern was identifiable three to four days before the actual event on the upper-air weather maps, on May 28. A short-wave trough within the well-defined North Atlantic frontal zone moved southeastward from Newfoundland in Canada, gradually increasing in amplitude. On May 30 the trough reached the west of the British Isles, absorbing an aged upper-level low near the Iberian Peninsula. West and Central Europe fell under a southwest flow, bringing a surge of subtropical warm air. This resulted in temperatures of over +18 °C at 850 hPa over southwestern Germany. The preceding dry period with low water supply meant the warm air's potential was fully utilized, leading to the first spring temperatures in western Germany exceeding +30 °C. Name, the town of Bendorf reached the highest temperature ever recorded for May with +33.7 °C. On May 31st the heat shifted to eastern Germany, with temperatures around Berlin reaching up to +34 °C (Wettergefahren-Frühwarnung, Insitut für Meteorologie und Klimaforschung, KIT).

My model output for this time period (Figure 4.11) shows that precipitation fall starts appearing on the morning of May 31st in Western Germany (Fig. 4.11a). Later in the afternoon, the field moves toward the east, notably reaching over 8 mm of rain in areas of the North and South Germany (Fig. 4.11b). On June 1st the storm system starts leaving the area (Fig. 4.11c), while in the afternoon and evening hours scattered rainfall appears in

the southwest (Fig. 4.11d). Generally, my model output aligns very well with the Early Warning Report and its analysis of the event; it captures the rainfall amounts in the north and south correctly, and it reflects the eastward movement of the storm and the heavy precipitation in the reported areas. Additionally, the model's depiction of scattered rainfall resuming in the southwest on June 1st aligns with the report's description of the post-frontal situation.



(a) May 31st 2011, 09:00 CET

(b) May 31st 2011, 17:00 CET

(c) June 1st 2011, 06:00 CET

(d) June 1st 2011, 20:00 CET

Figure 4.11: Plots of precipitation fields for selected timesteps during the events on May/June 2011: (a) May 31st, 09:00, (b) May 31st, 17:00, (c) June 1st, 06:00, and (d) June 1st, 20:00

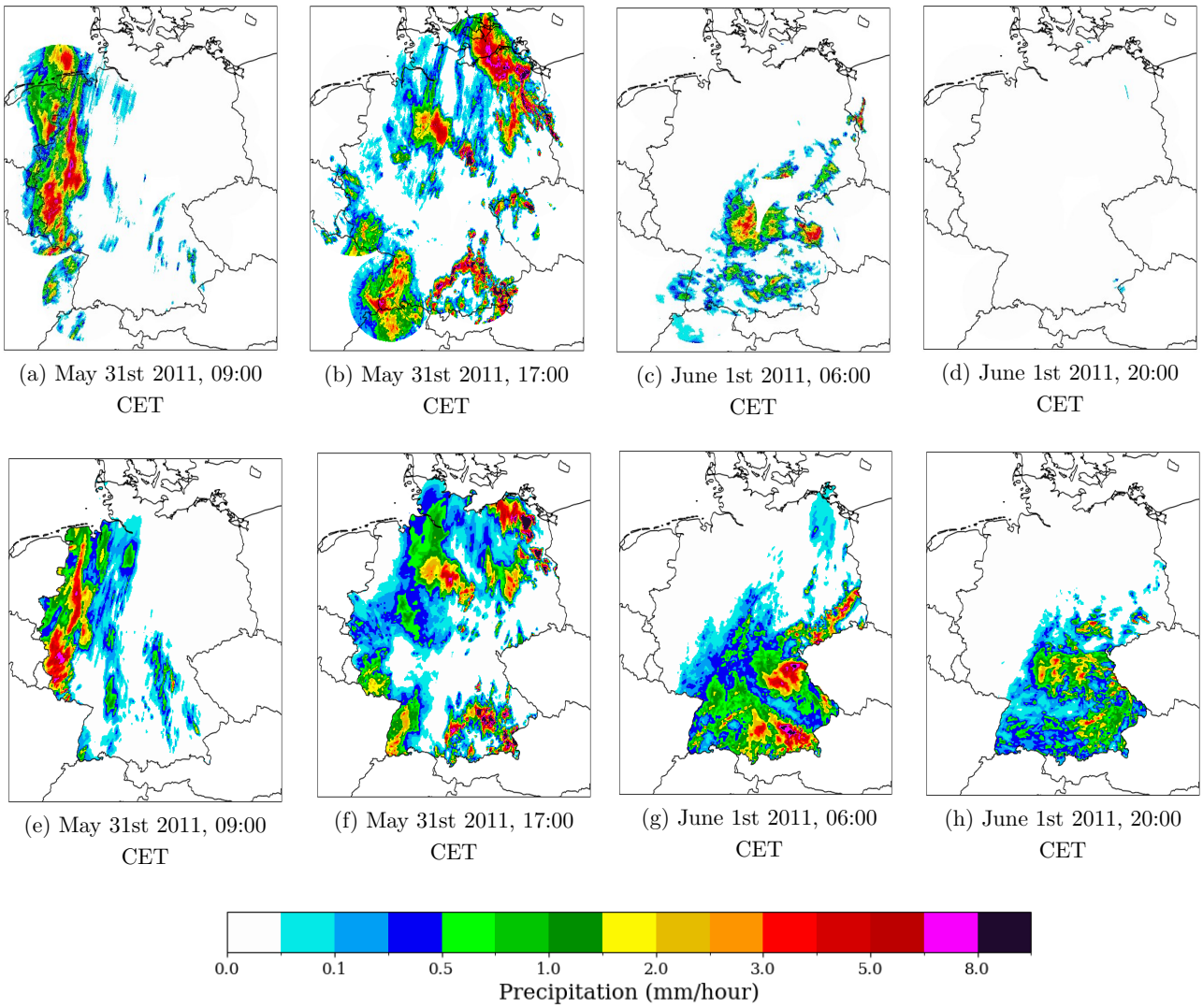


Figure 4.12: Comparative precipitation intensity visualizations from two datasets at selected time steps during the weather events spanning May to June 2011. The first row depicts RADKLIM data for (a) May 31st, 09:00, (b) May 31st, 17:00, (c) June 1st, 06:00, and (d) June 1st, 20:00. The second row presents REA2 dataset visualizations for corresponding times (e) through (h).

As shown in Figure 4.12, similar conclusions for the development of the event can be drawn from REA2 and RADKLIM since they both show a progression of the precipitation event over time, with the area of precipitation changing in shape and intensity. The model output also shows changes over time, but the evolution might differ in terms of where the most significant changes in precipitation occur. Another difference found in the output is related to the spatial distribution of precipitation. The model seems to reconstruct more concentrated areas of high precipitation, as seen in 4.11(a) and 4.11(b). It is nevertheless able to capture small-scale convective events. When investigating 4.11(c) and 4.12(c) I notice that in the middle of the RADKLIM pattern, there is a missing radar pattern, possibly due to radar malfunction, as discussed in the previous section. The unavailability

of radar data again poses a problem here, as it coincides with a precipitation event occurring precisely within the region affected by the malfunctioning radar. In contrast, my model effectively mitigates this issue by providing data infilling, thereby enabling us to analyze the precipitation event. Interestingly, on the final timestep of Figures 4.11 and 4.12 on 21:00 I notice that my model output and RADKLIM show very little precipitation fall (less than 3 mm and within a very small region in the southeast), whereas REA2 shows multiple large scale events occurring everywhere in South Germany. These extensive patterns present in the reanalysis data are not reported on the records by the DWD or the Early Warning Report published by KIT and as such, they do not seem close to reality. Another drawback of the REA2 representations is that there is significant misplacement of events, for example when comparing (a) and (c) of Figure 4.12 with (e) and (g) respectively. This is a peculiar result since REA2 integrates radar data in its data assimilation scheme (Wahl et al. 2017), so it should be more in agreement with RADKLIM. The reported discrepancies can be attributed to the differences in grid and perhaps in the method the hourly precipitation sums are produced in each dataset.

Apart from the investigation of the total precipitation field, I further use a point-wise comparison, where I calculate the total precipitation sums over 24 hours for my model output and REA2 (Table 4.1). Consecutively I compare my sums to the amounts measured at specific stations, as stated in the report. This is a particularly difficult task, as it requires for the gridded datasets not only to have captured the right amount of precipitation, but place it in the exact same location as the station as well. The stations are located in Bavaria, since this is the region where the event I am investigating took place. My model's output significantly underestimates the precipitation sums, and in four of the five stations it is even half of the actual measured amount. Interestingly, the most accurately-captured amount is found in Kümmersbruck (difference of 23 mm), which is not far from Schmidmühlen where I had an underestimation of 37 mm, while the target in both locations was the same (58 mm). It is not clear why this difference is present, since the two stations are relatively close and there are no major differences in the landscape.

Station	Precipitation Sum 31/05 – 01/06 (mm)	
	Model Output	KIT Report
Stammbach-Querenbach	17	66
Kümmersbruck	35	58
Schmidmühlen	21	58
Freystadt-Oberndorf	20	52
Waakirchen-Demmelberg	18	51

Table 4.1: Comparison of precipitation sums over five station locations. The sums were calculated for a 24 hour duration on May 31st – June 1st at 06:00, in the year 2011.

Next, I investigate monthly mean time series. It is crucial to acknowledge that while analyzing trends and variations in precipitation data, one could consider precipitation sums. However, given the variation in spatial resolution among data sources, this approach may not be suitable for my comparative analysis. The reference datasets (ERA5, RADKLIM, REA2) are gridded with differing grid sizes, whereas the station data are found as daily values in pairs of latitude and longitude. Comparing precipitation sums from a limited number of station points (potentially a few hundred) to those from a significantly larger number of gridded points would be inappropriate. For instance, the REA2 dataset contains about 500,000 points, ERA5 has 1,890 points, and RADKLIM and my model output encompass 990,000 points each. Consequently, the sums derived from these gridded fields would be substantially larger and would not offer a meaningful comparison to the observational data from the stations.

Figure 4.13 serves as a comparison of monthly mean precipitation rates across various sources for the year 2011; I utilize three reference datasets, REA2 , RADKLIM and ERA5. I also calculate the monthly mean precipitation rates as given in stations that measure at a daily temporal resolution. The number of these stations is close to 700. All datasets show a similar seasonal pattern, with the lowest mean precipitation in the earlier months (January to March) and higher values in the middle of the year (May to August). Moreover, all datasets indicate that the peak precipitation occurs in July. The daily stations and RADKLIM agree significantly, which is an expected outcome since RADKLIM uses station measurements to a great extent. There is a significant divergence among all datasets in June and July, which are among the months with the highest precipitation. This could be attributed to different systems capturing different aspects of what might be a highly variable precipitation event during these months. Furthermore, summer precipitation, as it is often caused by convection, it manifests itself through sudden and

intense events that can be challenging to capture in models. Stations, on the other hand, due to their direct measurement methodology, might be more succesful in capturing such events. This could explain why the station monthly means reach high values in the summer months. The underestimation in precipitation fall from my model is evident on this plot, similar to Figures 4.6 and 4.8. The primary challenge for the model in capturing convective processes contributes to a more pronounced bias, particularly during the summer months. Additionally, accurately predicting sudden and intense weather events, which occur not only in summer but throughout the year, poses a significant challenge for the model. This difficulty arises not only from the abrupt nature of these events but also due to the scarcity of extreme events in the model's training dataset.

An interesting observation is that ERA5 overestimates precipitation greatly, marking the highest monthly mean values for most of the year. According to Hu et al. (2020) ERA5 has certain limitations in accurately capturing precipitation extremes. Moreover, the authors mention that the precipitation in ERA5 is generated by a combination of large-scale cloud and precipitation schemes and a convection scheme. These modeling choices and the nature of the data assimilation system could contribute to discrepancies in precipitation representation. In another recent work, Lavers et al. (2022) discuss that the wet bias noted in ERA5 in their global analysis can be attributed to differences in elevation and orography. This argument is especially fitting in my result, since I detect a constant wet bias especially when focusing on the ERA5 monthly means compared to the monthly means derived from daily stations.

As far as REA2 is concerned, I see that it has strong agreement with RADKLIM, for the majority of months. The reason for this is most likely the data assimilation scheme used in REA2, which utilizes the same radar scans included in RADKLIM. Wahl et al. (2017) report that for the whole range 2007 – 2013 this dataset tends to overestimate spring and summer precipitation. REA2, unlike its predecessors, does not include a parametrization of deep convection, and instead assumes that phenomena on the convective scale can be resolved by the model scale. Thus, discrepancies in observed precipitation can stem from errors in this setup, that can be evident even if I am focusing only on the year 2011, unlike Wahl et al. that studied the full time span of the dataset.

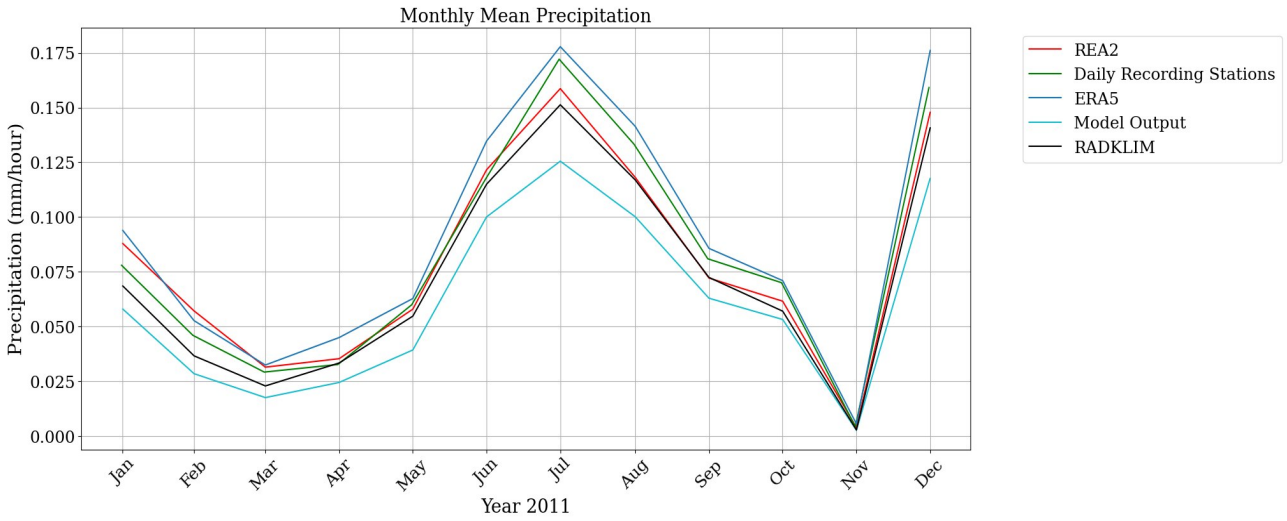


Figure 4.13: Comparison of monthly mean precipitation for the year 2011 as recorded by different datasets: the AI model output (cyan), REA2 (red), RADKLIM (black), observations from ground stations that operate daily (green), and ERA5 (blue).

4.3.3 The 1999 Pentecost flood

The 1999 Pentecost flood, also known as "Pfingsthochwasser 1999," was a devastating flood event that struck in May 1999, predominantly impacting Bavaria in Germany, and Vorarlberg and Tirol in Austria. Fuchs et al. (1999) characterized it as a centennial occurrence due to its severe impact. Similar to the summer floods of 2021, the 1999 event resulted from a combination of factors rather than just precipitation during the flood days. While the meteorological conditions of May 1999 were not extraordinary, the convergence of continuous heavy rain, melting snow, and pre-saturated soils led to this catastrophic event. The floods were preceded by several months of intense precipitation in the Alpine region. Notably, from February 1999, substantial snow accumulation began, with a remarkable increase of 160 cm in snow cover on the Zugspitze mountain over just seven days (February 17 to 24). In April, precipitation levels were significantly above average, leading to highly moist soil conditions. This combination of elevated soil moisture and substantial snow cover created the ideal conditions for the occurrence of the May floods.

On the 20th of May, a low-pressure area extended from the Adriatic Sea to Northeast Germany, causing moist warm air from the Balkans and moist cool air from the Atlantic to meet and create a frontal system. This situation was exacerbated by topographical features like the Alps, which forced the air upwards, enhancing the formation of precipitation. The result was several days of heavy rainfall in southern

Germany and the northern Alps. In Figure 4.14 I compare the depiction of a snapshot during the event taken from different sources, for the area of intense precipitation in the Bavarian Alps. It is important to note that the snapshot shown in Fig. 4.14c, acquired from Fuchs et al. (1999) utilizes only 15 radars, since the 17 radars that comprise the full DWD network nowadays, were not yet in place. Moreover, given the heavily pixelated scan quality, I infer that these are not double-polarization radars, which suggests potential inaccuracies in their measurements beyond the issue of pixelated reflectivity scans. While my model may slightly underestimate the extent of precipitation toward central Germany, it effectively captures the areas of intense rainfall in the south. Considering the challenging topography and mountainous regions in this area, it's noteworthy that my model represents it accurately. Additionally, the absence of checkerboard artifacts in the remainder of the map is a positive aspect. In comparison, ERA5 also manages to capture the southern event despite its significantly coarser resolution. However, it is important to highlight that ERA5 tends to overestimate precipitation in other parts of the country, possibly due to elevation differences as discussed in Section 4.3.2. This overestimation is a critical point of difference between the two models.

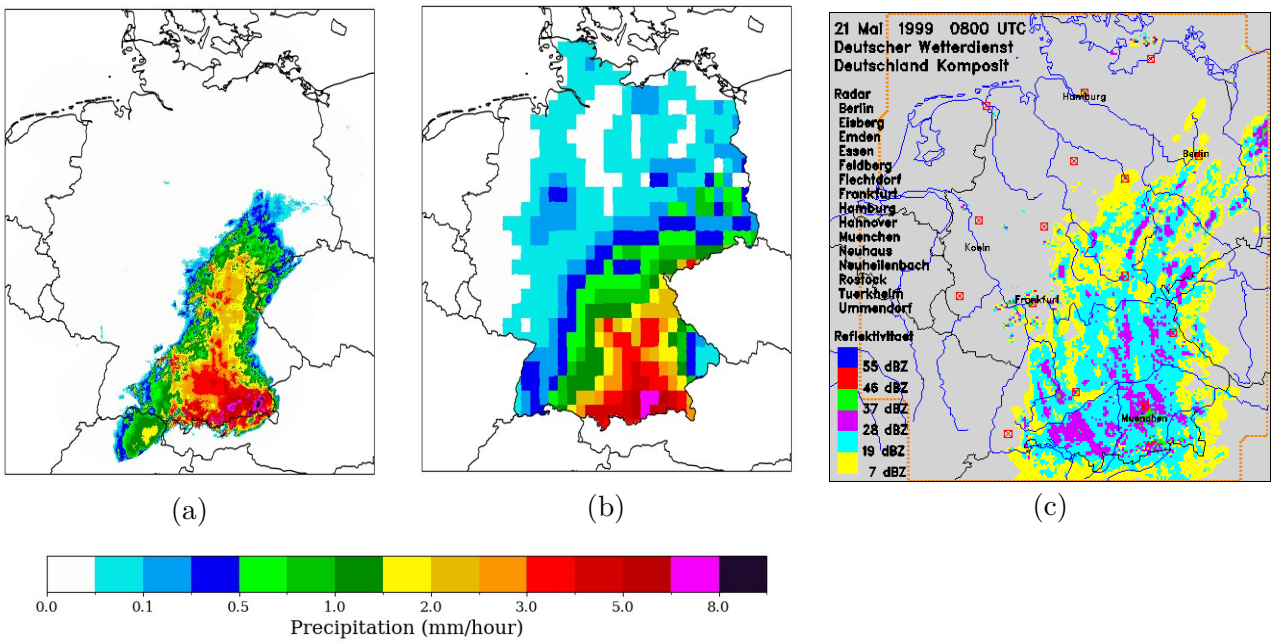
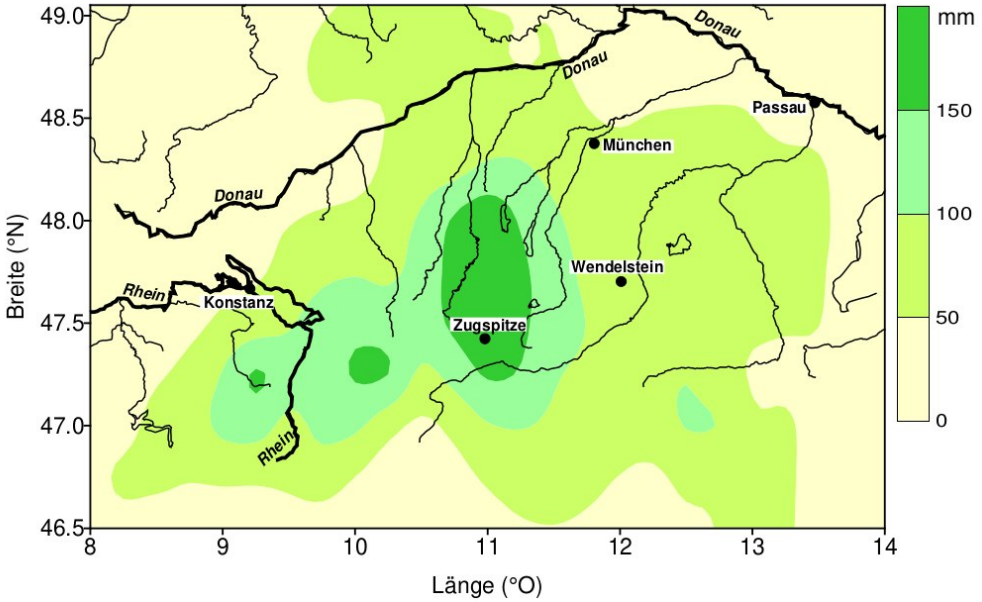


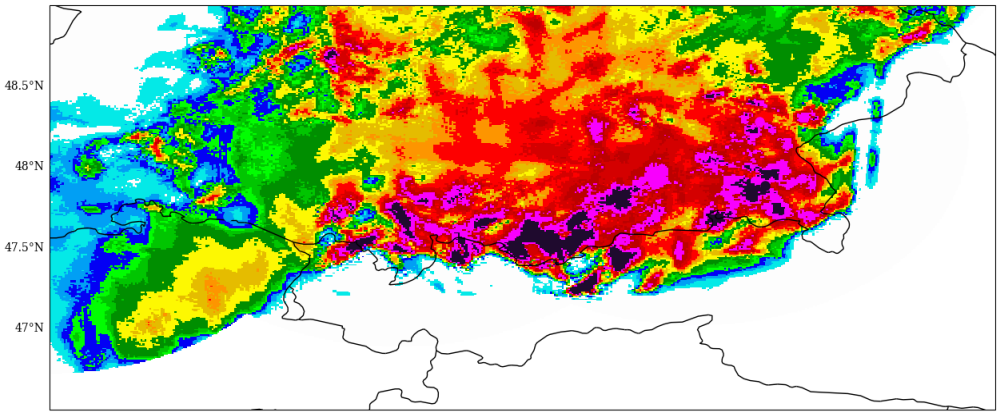
Figure 4.14: Depiction of the timestep 21st May,1999 at 10:00 across three datasets: (a) the AI model output, (b) ERA5 reanalysis and (c) reflectivity of DWD radar scans (Fuchs et al. 1999).

In the report published by the DWD shortly after the event, data from 400 precipitation stations of varying temporal resolution (hourly, daily and monthly) were used at the Alpine region (Fuchs et al. 1999). The data were corrected using a method to account for wind and evaporation influence, and they were interpolated using kriging. The results are depicted in Fig. 4.15c. I recreate the same area with the precipitation sums for three days spanning 20 - 22 May 1999 derived from my model output (Fig. 4.15a) and ERA5 (Fig. 4.15b). Despite my model using only 8 points in that region (Fig. 4.15d) I notice that the high intensity area around Zugspitze is very well captured. It also has a more significant spread across the southern border of Germany, compared to the DWD plot. This spread also appears in ERA5, though due to that dataset's coarser resolution the high intensity covers an even bigger area. In contrast to the DWD report and ERA5, my model captures little to no rainfall in the area near Passau. The most likely explanation for this issue is that the model lacks station information near the affected area, making accurate predictions challenging. Supporting this hypothesis are the time correlation plots in Figures 3.7 and 3.8, which demonstrate that this particular area typically exhibits lower correlations compared to the reference datasets HYRAS in 1997 and RADKLIM in 2018.

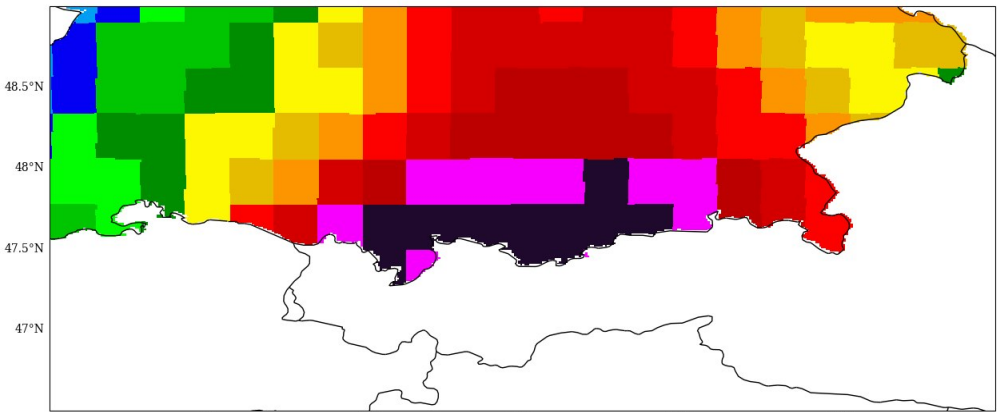
Figure 4.16, similarly to Fig. 4.13, depicts the monthly mean precipitation of ERA5, my model output and 400 daily operating stations for the year 1999. Both datasets accurately capture the annual and seasonal patterns. As observed before, compared to the stations, ERA5 has a wet bias and my model has a dry bias. However, the wet bias in this year for ERA5 is not as strong as in 2011, in contract to my output, where the bias remains in the same levels as before. A plausible explanation for the reduced bias in ERA5 could be attributed to the use of almost half the number of stations compared to previous analyses. As previously discussed, the bias in ERA5 is likely due to elevation differences at the station locations. With fewer stations, the impact of this elevation discrepancy is diminished. Notably, in this particular year, ERA5 more effectively captures the highest precipitation values. Conversely, my model does not demonstrate the same level of skill in as observed in the 2011 plot, as far as monthly means are concerned. The model struggles to replicate the sharp fluctuations shown by the station data, instead producing a more smoothed-out plot. While it successfully identifies the peaks and troughs, these are not as distinct as they ought to be. This highlights the need for additional refinement in the model's capacity to accurately capture extreme variations.


Donau
Donau
Donau
Rhein
Rhein
Passau
München
Wendelstein
Zugspitze
Konstanz
mm
150
100
50
0
49.0
48.5
48.0
47.5
47.0
46.5
Breite (°N)
8
9
10
11
12
13
14
Länge (°O)

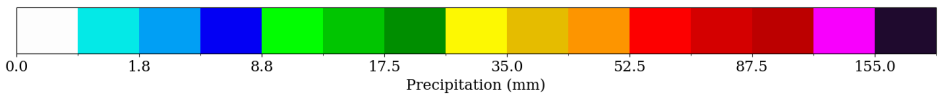
(a)

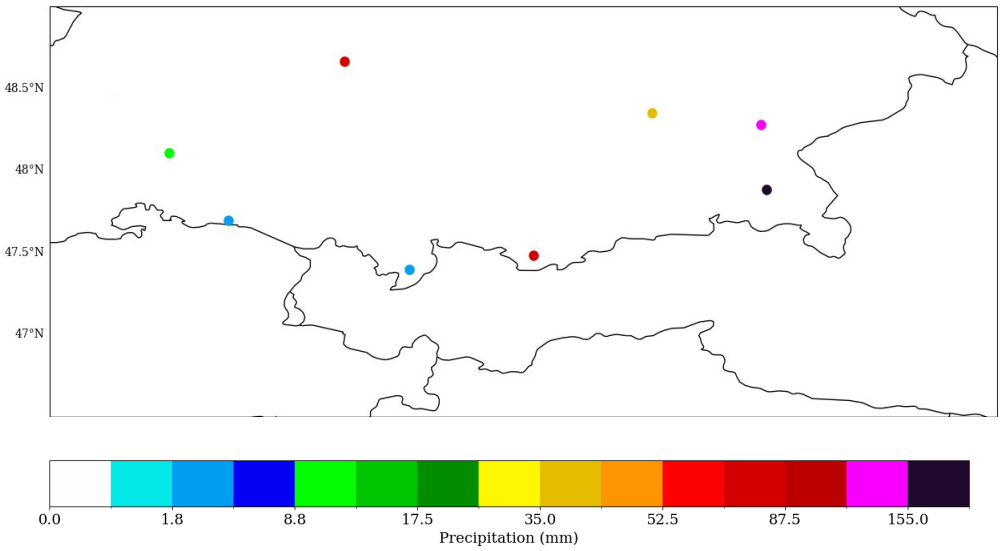

48.5°N
48°N
47.5°N
47°N

(b)


48.5°N
48°N
47.5°N
47°N

(c)


0.0
1.8
8.8
17.5
35.0
52.5
87.5
155.0
Precipitation (mm)



(d)

Figure 4.15: Precipitation totals from May 20 to 22, 1999, comparing (a) interpolated DWD report stations (Fuchs et al. 1999), (b) AI model's output, (c) ERA5, and (d) operational hourly basis stations—the sole points utilized by the AI model in the specified region.

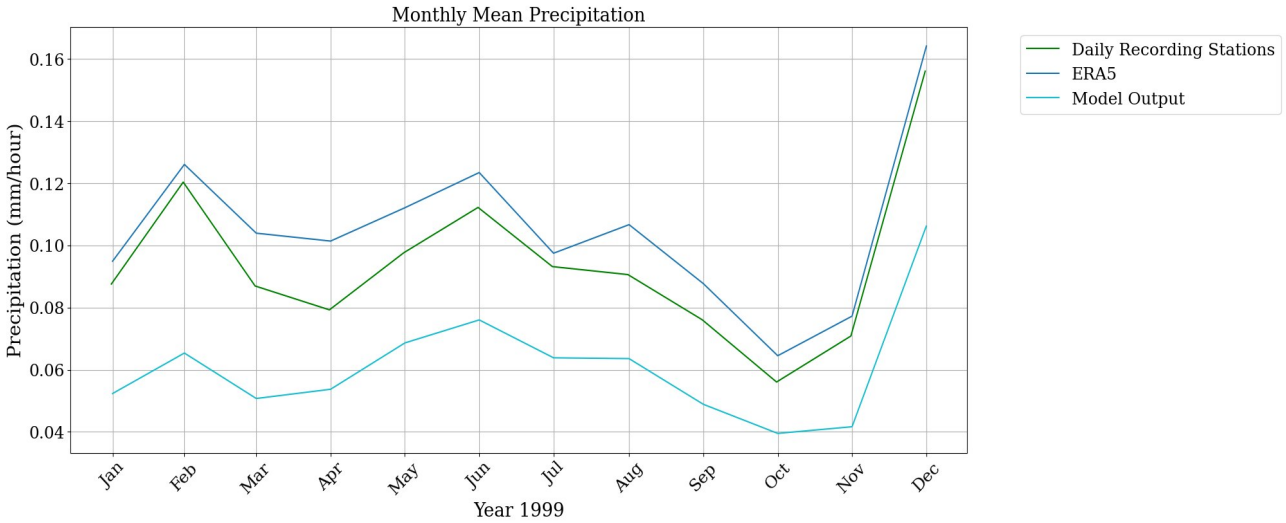


Figure 4.16: Comparison of monthly mean precipitation for the year 1999 as recorded by different datasets: the AI model output (cyan), observations from ground stations that operate daily (green), and ERA5 (blue).

An important aspect that needs to be discussed is possible implications of the monthly averaging technique. It is important to note that in both Figures 4.13 and 4.16, datasets with different spatial resolutions are compared. Particularly in Fig. 4.16, I compare the monthly field averages of ERA5 (on a 54 x 35 grid), the much more detailed

model output (900 x 1100 grid) and the station averages, which are mere data points on the map, not gridded products. The coarser grid of ERA5 might reduce the 'noise' in the data, leading to less internal variations within the same event, at the expense of losing finer details (like convective scale characteristics). In contrast, my more detailed model output might encompass more noise, as indicated by the checkerboard artifacts, and reveal more nuanced variations within the same event. This increased detail could contribute to a lower overall average. Additionally, the scarcity of extreme events in the training dataset is another factor that might explain why my model tends to underestimate monthly mean precipitation amounts.

Next, similarly to Table 4.1, I conduct a comparison between actual measurements from daily stations and the daily cumulative values from my hourly reconstructed precipitation field (Table 4.2). This comparison is based on daily measurements at the stations and my calculated daily totals for the entire day of May 21st, 1999. The results show a significant agreement. In key areas like Zugspitze and Wendelstein, the model slightly overestimates precipitation by a few millimeters, but this deviation falls within an acceptable range. Conversely, at three other locations - Weissach near Stuttgart, Munich, and the Black Forest - there is a slight underestimation. Notably, Weissach did not record major precipitation amounts, aligning with this trend. These findings agree with the precipitation field depicted in Figure 4.15, as well as with the associated DWD report. The areas where overestimation occurs are characterized by mountainous terrain in the South, near the Alps, posing a challenge for accurate depiction due to their involvement in strong convective scale events. Despite these difficulties, the model's close accuracy, missing by only a few millimeters, is commendable. The other three locations, being farther away and situated in areas without intense convection and on flatlands, might seem easier to predict. However, their distance from the stations used in the model input can affect the model's accuracy. Despite these challenges, and using only 76 data points to reconstruct this detailed convective event, the model's performance is notable, particularly considering that is used information from only 8 stations in that region.

Station	Station Measurement	Model Output
Zugspitze	133.68	134.91
Weissach	2	0.98
Wendelstein	22.08	32.73
Schoenwald - Schwarzwald	2.8	1.98
Munich	50.16	44.95

Table 4.2: Comparison of precipitation daily sums (May 21st, 1999) over five station locations. The names of the stations are given in the first column, the measurement of each station is given in the second column, while the third column shows the measurement taken from the AI model output in the grid cell closest to that station. These 5 stations were not among the 76 stations that were used as input for the model.

5. Conclusion and Outlook

In this work, I applied a novel Machine Learning method to reconstruct high spatially resolved precipitation fields over Germany, at an hourly temporal resolution. The goal of the thesis was to enhance the representation of precipitation in the past at a regional scale. I used precipitation data from different sources (stations, radars and reanalysis) to organize a number of experiments that utilized various training methodologies. The different training configurations incorporated a series of meteorological variables, as well as temporal information in the form of multiple input timesteps for training. I compared the experiments by employing several key metrics, including: the root mean square error (RMSE), the field and time correlation coefficients and the difference of total precipitation. The metric calculations were done between my models outputs and the reference datasets HYRAS and RADKLIM.

The models presented diverse levels of accuracy across multiple metrics. Notably, those integrating station wind measurements or incorporating multiple timesteps for prediction consistently yielded lower RMSE. All models demonstrated moderate field correlation coefficients. Additionally, the time correlation coefficients on grid points indicated strong correlations in most regions, highlighting consistent model performance over time across Germany. However, a common trend among the models was the tendency to underestimate total precipitation, particularly during high-intensity events.

Based on the evaluated metrics, the most effective model was identified as the one incorporating multiple input timesteps for prediction. Utilizing this model, precipitation data was reconstructed for the period spanning 1995 to 2000 when no highly-resolved radar dataset are available. The reconstructed data exhibited several noteworthy characteristics, including reasonable agreement with reference datasets in both spatial and temporal precipitation patterns. However, there were instances of underestimation and misplacement of precipitation events, particularly noticeable during periods of intense precipitation. Furthermore, a consistent dry bias was observed in comparison to corresponding reference data from HYRAS and RADKLIM.

I performed an event-based approach, where I chose events along different years (2021, 2011 and 1999). For all the events my model produced plausible results which were comparable to recent high-end datasets. I effectively used the 2021 summer flood as a benchmark, showcasing that my model closely aligns with RADKLIM data. Although there were minor discrepancies in intensity and distribution, the model accurately captured the general patterns and trends of these severe weather events, including a detailed analysis of 24-hour and 72-hour precipitation sums. For a significant cold front event in May 2011, the model's output agreed with the actual meteorological developments, capturing the onset and eastward movement of the storm. The model accurately represented rainfall amounts and small-scale convective events, despite underestimating precipitation sums at specific station locations. The model's reconstruction of the 1999 Pentecost flood was particularly impressive. It successfully captured the complex convective event driven by heavy rain, melting snow, and pre-saturated soils, particularly in the challenging topography of southern Germany. Compared to outdated radar data and sparse station measurements prevalent in the area, our model excelled in reconstructing a comprehensive precipitation field that effectively surpassed existing data sources. Furthermore, when compared to the reanalysis dataset ERA5, which requires considerable computational resources to produce, our model achieved better event representation due to higher resolution.

It is evident that while the AI model marks significant advancements, there remains a series of opportunities for further improvement. Several results highlighted the importance of temporal information in predicting precipitation. Previous works (Meuer et al. 2022) proved that incorporating Long Short-Term Memory (LSTM) networks in the baseline model used here substantially enhances the model's performance. Another strategy could include employing more granular data, like 5-minute timesteps with the goal of refining predictions of short-term variations. While acknowledging that some variables, such as wind and pressure, provided insights on specific cases, it should be noted that to fully take advantage of these additional variables, modifications in the model's architecture are essential. This may involve incorporating more layers and using a deeper model. Moreover, addressing the discrepancies caused by placing station data on the RADKLIM grid could lead to more accurate spatial representations. Correcting for these grid

differences would ensure that the model better aligns with the actual geographic distribution of weather phenomena. A limitation in my work is found in the type of analysis that was conducted. I applied an event-based evaluation on my model and did not conduct more detailed statistical analysis, in the case of extreme events and climate trends. In future works the interpretation of the results could benefit from such analyses. A challenge in the time series analysis of my reconstructions has been a constant dry bias. Leveraging this bias as a corrective factor could significantly improve the accuracy of the reconstructed fields, aligning the model outputs more closely with observed data trends. Furthermore, refining the loss functions, potentially by tweaking the loss weights, could effectively address the issue of checkerboard artifacts observed in the model's output, a phenomenon also reported in prior studies such as those by Liu et al. (2018). Finally, considering the model's training process, a strategic refinement of the training set is proposed. By selectively excluding timesteps that feature low intensity precipitation events and maintaining a balance between heavy rain and dry periods, the model can be trained more equitably across diverse weather conditions.

In retrospect, it is crucial to acknowledge the model's commendable performance in reconstructing the past, particularly given the limited data points (76 points) used to reconstruct a substantially larger dataset (990,000 points). The model adeptly captured complex patterns of convection and extreme weather events, despite not utilizing extensive quality assessment, denoising, and numerical corrections typically employed in climate dataset processing. In conclusion, the AI model's successful performance suggests promising prospects for broader temporal applications beyond the 1995 threshold, leveraging the availability of relevant data of the past. The outlined future perspectives provide a roadmap for advancing its capabilities further in climate and precipitation analysis.

References

[1] Junghänel, T.; Bissolli, P.; Daßler, J.; Fleckenstein, R.; Imbery, F.; Janssen, W.; Kaspar, F.; Lengfeld, K.; Leppelt, T.; Rauthe, M.; et al. Hydro-klimatologische Einordnung der Stark- und Dauerniederschläge in Teilen Deutschlands im Zusammenhang mit dem Tiefdruckgebiet "Bernd" Vom 12. Bis 19. Juli 2021; DWD: Offenbach, Germany, 2021.

[2] Müller C.; Nied, M.; Voigt. M; Iber, C; Sauer, T.; Junghaänel, T; Hoy, A; Hübener, H; Starkniederschläge - Entwicklungen in Vergangenheit und Zukunft, Kooperation KLIWA, Germany, 2019

[3] Mauch, F., "The Great Flood of 1962 in Hamburg." Environment & Society Portal, *Arcadia*, no. 6. Rachel Carson Center for Environment and Society, München, Germany, 2012

[4] Cristiano, Elena, Marie-Claire ten Veldhuis, and Nick van de Giesen. "Spatial and Temporal Variability of Rainfall and Their Effects on Hydrological Response in Urban Areas – A Review." *Hydrology and Earth System Sciences* 21, no. 7 (2017): 3859–78

[5] Deutscher Wetterdienst - Data Assimilation. Accessed October 4, 2023. https://www.dwd.de/EN/research/weatherforecasting/num_modelling/02_data_assimilation/data_assimilation_node.html.

[6] Yin, Shuiqing, and Deliang Chen. "Weather Generators." *Oxford Research Encyclopedia of Climate Science*, 2020.

[7] Ailliot, Pierre; Allard, Denis; Monbet, Valérie; Naveau, Philippe. Stochastic weather generators: an overview of weather type models. Journal de la société française de statistique, Volume 156 (2015) no. 1, pp. 101-113.

[8] Richardson, C. W. "Stochastic Simulation of Daily Precipitation, Temperature, and Solar Radiation." *Water Resources Research* 17, no. 1 (1981): 182–90.

[9] Maraun, Douglas and Martin Windmann. *Statistical Downscaling and Bias Correction for Climate Research.* "Chapter 13 - Weather Generators." s.l.: Cambridge University Press, 2018.

[10] Furrer, Eva M., and Richard W. Katz. "Improving the Simulation of Extreme Precipitation Events by Stochastic Weather Generators." *Water Resources Research* 44, no. 12 (2008).

[11] Yang, C., R. E. Chandler, V. S. Isham, and H. S. Wheater. "Spatial-Temporal Rainfall Simulation Using Generalized Linear Models." *Water Resources Research* 41, no. 11 (2005).

[12] New, M, D Lister, M Hulme, and I Makin. “A High-Resolution Data Set of Surface Climate over Global Land Areas.” *Climate Research* 21 (2002): 1–25.

[13] Oliver, M. A. and R. Webster “Kriging: A Method of Interpolation for Geographical Information Systems.” *International journal of geographical information systems* 4, no. 3 (1990): 313–32.

[14] Auchincloss, Amy H., Ana V. Diez Roux, Daniel G. Brown, Trivellore E. Raghunathan, and Christine A. Erdmann. “Filling the Gaps: Spatial Interpolation of Residential Survey Data in the Estimation of Neighborhood Characteristics.” *Epidemiology* 18, no. 4 (2007): 469–78.

[15] Haberlandt, Uwe. “Geostatistical Interpolation of Hourly Precipitation from Rain Gauges and Radar for a Large-Scale Extreme Rainfall Event.” *Journal of Hydrology* 332, no. 1–2 (2007): 144–57.

[16] Kurth, Thorsten, Shashank Subramanian, Peter Harrington, Jaideep Pathak, Morteza Mardani, David Hall, Andrea Miele, Karthik Kashinath, and Anima Anandkumar. “FourCastNet: Accelerating Global High-Resolution Weather Forecasting Using Adaptive Fourier Neural Operators.” *Proceedings of the Platform for Advanced Scientific Computing Conference*, 2023.

[17] Kadow, Christopher ; Hall, David M. ; Ulbrich, Uwe: Artificial intelligence reconstructs missing climate information, Nature Geoscience 13 (2020), Nr. 6, S. 408–413

[18] Liu, Guilin ; Reda, Fitsum A. ; SHIH, Kevin J. ; Wang, Ting-Chun ; Tao, Andrew ; Catanzaro, Bryan: Image inpainting for irregular holes using partial convolutions, Proceedings of the European Conference on Computer Vision (ECCV), 2018, S. 85–100

[19] Bárdossy, András, and Geoffrey Pegram. “Infilling Missing Precipitation Records – a Comparison of a New Copula-Based Method with Other Techniques.” *Journal of Hydrology* 519 (2014): 1162–70

[20] Verworn, A., and U. Haberlandt. “Spatial Interpolation of Hourly Rainfall – Effect of Additional Information, Variogram Inference and Storm Properties.” *Hydrology and Earth System Sciences* 15, no. 2 (2011): 569–84

[21] Berndt, C., and U. Haberlandt. “Spatial Interpolation of Climate Variables in Northern Germany—Influence of Temporal Resolution and Network Density.” *Journal of Hydrology: Regional Studies* 15 (2018): 184–202

[22] Londhe, Shreenivas, Pradnya Dixit, Shalaka Shah, and Shweta Narkhede. “Infilling of Missing Daily Rainfall Records Using Artificial Neural Network.” *ISH Journal of Hydraulic Engineering* 21, no. 3 (2015): 255–64

[23] Coulibaly, P., and N.D. Evora. “Comparison of Neural Network Methods for Infilling Missing Daily Weather Records.” *Journal of Hydrology* 341, no. 1–2 (2007): 27–41

[24] Militino, Ana F., María Dolores Ugarte, and Unai Pérez-Goya. “Machine Learning Procedures for Daily Interpolation of Rainfall in Navarre (Spain).” *Trends in Mathematical, Information and Data Sciences*, 2022, 399–413

[25] Moraux, Arthur, Steven Dewitte, Bruno Cornelis, and Adrian Munteanu. "Deep Learning for Precipitation Estimation from Satellite and Rain Gauges Measurements." *Remote Sensing* 11, no. 21 (2019): 2463

[26] Rojas-Campos, Adrian, Michael Langguth, Martin Wittenbrink, and Gordon Pipa. *Deep learning models for generation of precipitation maps based on numerical weather prediction*, 2022

[27] Teegavarapu, Ramesh S., Alaa Aly, Chandra S. Pathak, Jon Ahlquist, Henry Fuelberg, and Jill Hood. "Infilling Missing Precipitation Records Using Variants of Spatial Interpolation and Data-driven Methods: Use of Optimal Weighting Parameters and Nearest Neighbour-based Corrections." *International Journal of Climatology* 38, no. 2 (2017): 776–93

[28] Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: Patchmatch: A randomized correspondence algorithm for structural image editing. ACM Transactions on Graphics-TOG 28(3), 24 (2009)

[29] Cai, Nian, Zhenghang Su, Zhineng Lin, Han Wang, Zhijing Yang, and Bingo Wing-Kuen Ling. "Blind Inpainting Using the Fully Convolutional Neural Network." *The Visual Computer* 33, no. 2 (2015): 249–61

[30] Yu, Jiahui, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. "Generative Image Inpainting with Contextual Attention." *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018

[31] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) 115(3), 211–252 (2015).

[32] Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)

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

[34] Compo, Gilbert P. , Whitaker, Jeffrey S. , Sardeshmukh, Prashant D., Matsui N., Allan, R., Yin, X., Gleason, Byron E., Vose, Russel S. , Rutledge, Glenn, Bessemoulin, P.: The twentieth century reanalysis project. In: Quarterly Journal of the Royal Meteorological Society 137 (2011), Nr. 654, S. 1–28

[35] Taylor, Karl E., Ronald J. Stouffer, and Gerald A. Meehl. "An Overview of CMIP5 and the Experiment Design." *Bulletin of the American Meteorological Society* 93, no. 4 (2012): 485–98.

[36] Morice, Colin P., John J. Kennedy, Nick A. Rayner, and Phil D. Jones. "Quantifying Uncertainties in Global and Regional Temperature Change Using an Ensemble of Observational Estimates: The Hadcrut4 Data Set." *Journal of Geophysical Research: Atmospheres* 117, no. D8 (2012).

[37] Meuer, Johannes, “Infilling of Spatial Precipitation Recordings with a Memory-Assisted Convolutional Neural Network”, Master Thesis, Institut für Technische Informatik, Karlsruhe Institute of Technology (2022)
[38] “Radar-Online-Aneichung (RADOLAN).” Wetter und Klima – Deutscher Wetterdienst - Leistungen - Analysen radarbasierter stündlicher (RW) und täglicher (SF) Niederschlagshöhen. Accessed November 2, 2023.
https://www.dwd.de/DE/leistungen/radolan/radolan.htm.
[39] Winterrath, T., Brendel, C., Hafer, M., Junghänel, T., Klameth, A., Walawender, E., Weigl, E. and Becker, A., “Erstellung einer radargestuetzten Niederschlagsklimatologie”, Bericthe des Deutschen Wetterdienstes 251, Selbstverlag des Deutschen Wetterdienstes, Offenbach am Main (2017)
[40] “Radar Products.” Wetter und Klima - Deutscher Wetterdienst - Our services - Radar products. Accessed November 3, 2023. https://www.dwd.de/EN/ourservices/radar_products/radar_products.html.
[41] “Radar Network.” Wetter und Klima - Deutscher Wetterdienst - Our services - Radarverbund. Accessed November 3, 2023. https://www.dwd.de/DE/derdwd/messnetz/atmosphaerenbeobachtung/_functions/
[42] Kaspar, F., Müller-Westermeier, G., Penda, E., Mächel, H., Zimmermann, K., Kaiser-Weiss, A., & Deutschländer, T. (2013). Monitoring of climate change in Germany – data, products and services of Germany’s National Climate Data Centre. *Advances in Science and Research*, *10*(1), 99–106.
[43] Müller, C., Nied, M., Voigt, M., Sauer, T., Junghänel, T., Hoy, A., “Starkniederschläge Entwicklungen in Vergangenheit und Zukunft”, Kooperationsvorhaben KLIWA – Klimaveränderungen und Konsequenzen für die Wasserwirtschaft, Stand: 07/2019
[44] Kreklow J., Tetzlaff B., Kuhnt G., and Burkhard. B., “A Rainfall Data Intercomparison Dataset of Radklim, RADOLAN, and Rain Gauge Data for Germany.” *Data* 4, no. 3 (2019): 118.
[45] Kreklow, Jennifer, Björn Tetzlaff, Benjamin Burkhard, and Gerald Kuhnt. “Radar-Based Precipitation Climatology in Germany—Developments, Uncertainties and Potentials.” *Atmosphere* 11, no. 2 (2020): 217.
[46] Villarini, Gabriele, and Witold F. Krajewski. “Review of the Different Sources of Uncertainty in Single Polarization Radar-Based Estimates of Rainfall.” *Surveys in Geophysics* 31, no. 1 (2009): 107–29.
[47] Sauvageot, H. “Rainfall Measurement by Radar: A Review.” *Atmospheric Research* 35, no. 1 (1994): 27–54.
[48] Holleman, I., Michelson, D. B., Galli, G., Germann, U. and Peura, M. 2006. Quality information for radars and radar data: Deliverable: OPERA 2005 19. OPERA work package 1.2, 77 pp

[49] Bartels, H.; Weigl, E.; Reich, T.; Lang, W.; Wagner, A.; Kohler, O.; Gerlach, N. MeteoSolutions GmbH: Projekt RADOLAN—Routineverfahren zur Online-Aneichung der Radarniederschlagsdaten Mit Hilfe Von Automatischen Bodenniederschlagsstationen (Ombrometer); Zusammenfassender Abschlussbericht für die Projektlaufzeit von 1997 bis 2004; DWD: Offenbach, Germany, 2004.
[50] Ronneberger, Olaf ; Fischer, Philipp ; B Rox, Thomas: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention Springer, 2015, S. 234–241
[51] Simonyan, Karen ; Zisserman, Andrew: Very deep convolutional networks for large-scale image recognition. In: arXiv preprint arXiv:1409.1556 (2014)
[52] Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. In International Conference on Learning Representations.
[53] Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. arXiv preprint (2015)
[54] Rauthe, M., Steiner, H., Riediger, U., Mazurkiewicz, A., & Gratzki, A. (2013). A Central European precipitation climatology – part I: Generation and validation of a high-resolution gridded daily data set (HYRAS). *Meteorologische Zeitschrift*, *22*(3), 235–256.
[55] Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and Super-Resolution. *Computer Vision – ECCV 2016*, 694–711.
[56] Wahl, S., Bollmeyer, C., Crewell, S., Figura, C., Friederichs, P., Hense, A., Keller, J. D., & Ohlwein, C. (2017). A novel convective-scale regional reanalysis Cosmo-REA2: Improving the representation of precipitation. *Meteorologische Zeitschrift*, *26*(4), 345–361.
[57] "Late Spring Weather Event in Germany - May 2011." Wettergefahren-Frühwarnung, Insitut für Meteorologie und Klimaforschung, Karlsruhe Institute of Technology. Available at: https://www.wettergefahren-fruehwarnung.de/Ereignis/20110603_e.html.
[58] Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., … Thépaut, J. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049.
[59] Hu, G., & Franzke, C. L. (2020). Evaluation of daily precipitation extremes in reanalysis and gridded observation-based data sets over Germany. *Geophysical Research Letters*, *47*(18).
[60] Lavers, D. A., Simmons, A., Vamborg, F., & Rodwell, M. J. (2022). An evaluation of ERA5 precipitation for climate monitoring. *Quarterly Journal of the Royal Meteorological Society*, *148*(748), 3152–3165.
[61] Bandhauer, M., Isotta, F., Lakatos, M., Lussana, C., Båserud, L., Izsák, B., Szentes, O., Tveito, O. E., & Frei, C. (2021). Evaluation of daily precipitation analyses in e-obs (v19.0e) and era5 by comparison to regional high-resolution datasets in European regions. *International Journal of Climatology*, *42*(2), 727–747.

[62] Fuchs, T., Rapp. J. und B. Rudolf (1999): Starkniederschläge im Mai 1999 im Einzugsgebiet von Donau und Bodensee. Beilage Nr. 95/1999 zur Wetterkarte Nr. 158 und 159 des Deutschen Wetterdienstes. Selbstverlag, Offenbach am Main, Germany
[63] Bissolli,P; Göring L.; , Lefebvre Ch. (2021): Extreme Wetter – und Witterungsereignisse im 20. Jahrhundert; Deutscher Wetterdienst
[64] Lahoz, W. A., & Schneider, P. (2014). Data assimilation: making sense of Earth Observation. *Frontiers in Environmental Science, 2.*