spateGAN: Spatio-Temporal Downscaling of Rainfall Fields using a cGAN Approach

Luca Glawion[1], Julius Polz[1], Harald Günter Kunstmann[2], Benjamin Fersch[2], and Christian Chwala[1]

[1]Karlsruhe Institute of Technology
[2]Karlsruhe Institute of Technology (KIT)

February 20, 2023

Abstract

Climate models face limitations in their ability to accurately represent highly variable atmospheric phenomena. To resolve fine-scale physical processes, allowing for local impact assessments, downscaling techniques are essential. We propose spateGAN, a novel approach for spatio-temporal downscaling of precipitation data using conditional generative adversarial networks. Our method is based on a video super-resolution approach and trained on ten years of country wide radar observations for Germany. It simultaneously increases the spatial and temporal resolution of coarsened precipitation observations from 32 km to 2 km and from 1 hour to 10 minutes. Our experiments indicate that the ensembles of generated temporally consistent rainfall fields are in high agreement with the observational data. Spatial structures with plausible advection were accurately generated. Compared to trilinear interpolation and a classical convolutional neural network, the generative model reconstructs the resolution-dependent extreme value distribution with high skill. It showed a high Fractions Skill Score of 0.73 for rainfall intensities over 15mmh^{-1} and a low BIAS of 3.55%. A power spectrum analysis confirmed that the probabilistic downscaling ability of our model further increased its skill. We observed that neural network predictions may be interspersed by recurrent structures not related to rainfall climatology, which should be a known issue for future studies. We were able to mitigate them by using an appropriate model architecture and model selection process. Our findings suggest that spateGAN offers the potential to complement and further advance the development of climate model downscaling techniques, due to its performance and computational efficiency.

spateGAN: Spatio-Temporal Downscaling of Rainfall Fields using a cGAN Approach

Luca Glawion[1], Julius Polz[1], Harald Kunstmann[1,2], Benjamin Fersch[1], Christian Chwala[1,2]

[1]Institute of Meteorology and Climate Research, Karlsruhe Institute of Technology, Campus Alpin, Garmisch-Partenkirchen, Germany
[2]Chair of Regional Climate and Hydrology, Institute of Geography, University of Augsburg, Augsburg, Germany

Key Points:

- High performance simultaneous spatial and temporal precipitation downscaling enabled by 3D convolution approach
- Generation of realistic high-resolution ensembles using probabilistic conditional generative adversarial networks
- Low computational effort compared to dynamical downscaling approaches

Corresponding author: Luca Glawion, `luca.glawion@kit.edu`

Abstract

Climate models face limitations in their ability to accurately represent highly variable atmospheric phenomena. To resolve fine-scale physical processes, allowing for local impact assessments, downscaling techniques are essential. We propose spateGAN, a novel approach for spatio-temporal downscaling of precipitation data using conditional generative adversarial networks. Our method is based on a video super-resolution approach and trained on ten years of country wide radar observations for Germany. It simultaneously increases the spatial and temporal resolution of coarsened precipitation observations from 32 km to 2 km and from 1 hour to 10 minutes. Our experiments indicate that the ensembles of generated temporally consistent rainfall fields are in high agreement with the observational data. Spatial structures with plausible advection were accurately generated. Compared to trilinear interpolation and a classical convolutional neural network, the generative model reconstructs the resolution-dependent extreme value distribution with high skill. It showed a high Fractions Skill Score of 0.73 for rainfall intensities over $15\,mm\,h^{-1}$ and a low BIAS of 3.55%. A power spectrum analysis confirmed that the probabilistic downscaling ability of our model further increased its skill. We observed that neural network predictions may be interspersed by recurrent structures not related to rainfall climatology, which should be a known issue for future studies. We were able to mitigate them by using an appropriate model architecture and model selection process. Our findings suggest that spateGAN offers the potential to complement and further advance the development of climate model downscaling techniques, due to its performance and computational efficiency.

Plain Language Summary

Natural disasters like floods, hail, or landslides originate from precipitation. Global climate models are an important tool to understand these hazards and derive expected changes in a future climate. However, they operate on spatial and temporal scales that limit the regional ability to reflect their small scale characteristics. This has led to the development of dynamical and statistical downscaling methods. Due to their computational efficiency, machine learning algorithms recently get increased attention as method for improving the spatial resolution of climate data. Here, we describe a new deep learning model that allows to simultaneously increase both the temporal and spatial resolution of precipitation data. Our presented approach enhances the spatial resolution by a factor of 16 and the temporal resolution by factor of 6. The generated rain fields are hardly identifiable as artificial generated and exhibit the typical structure, movement and distribution of observed rain fields.

1 Introduction

In the 2010s around 83 % of all natural disasters were caused by weather and climate extremes killing more than 410,000 people. Half of all disasters were a direct consequence of precipitation extremes like floods or landslides (IFRC, 2021). Rising average temperatures are expected to further increase both mean and extreme precipitation (Seneviratne et al., 2021), a development that may even be underestimated in climate projections (Allan & Soden, 2008). In order to adapt to a changing climate, accurate local and global information about the current and future hydrological cycle is indispensable. However, precipitation shows high spatial and temporal variability, exhibiting fluctuations on almost all spatial and temporal scales (Berg et al., 2013). Dynamical global climate models are restricted to larger scales by their high computational demand and for numerical stability criteria. With typical horizontal grid spacing of 30–80 km (Chen et al., 2021) and temporal resolutions of 1–24 hours, they are beyond of resolving fine-scale physical processes, extreme precipitation in particular. Due to subgrid-scale parameterizations, conclusions about the development of small-scale processes under a changing climate are not generally limited. However, for physically-based local climate impact studies, the characterization of high-resolution information about precipitation and its extremes is inevitable.

Consequently, downscaling methods have been developed and applied to increase the resolution of climate model outputs. These methods include statistical and dynamical downscaling using regional climate models, as well as AI-based downscaling that leverages artificial neural networks (ANNs), which have become increasingly popular in recent years. The AI-based downscaling methods are based on the image "super-resolution" approach which originates from computer science, precisely computer vision, where the resolution of optical images is increased (Dong et al., 2016; Kim et al., 2016; J. Johnson et al., 2016). The logical extension of this approach to the temporal domain is called "video-super-resolution" (Lucas et al., 2018; X. Wang, Lucas, et al., 2019). While the original application of super-resolution is based on a clear understanding of the data-generating process, the processes of generating climate observations are less well understood, presenting both a challenge and an opportunity for the application of ANNs (Reichstein et al., 2019). Following the super-resolution approach, high-resolution observational, climate model, or reanalysis data are first spatially coarsened to a lower resolution. The training objective of the ANN is to recover the original resolution. For example, in precipitation downscaling, high-resolution weather radar observations enable the modeling of complex precipitation patterns using ANNs. An additional benefit of ANNs is a considerable reduction in computation time and energy compared to traditional dynamical models (Pathak et al., 2022).

First approaches for spatial precipitation downscaling with ANNs used a deterministic convolutional neural network (CNN) which does not account for potential biases between observations and global climate model data or cover uncertainties related to the highly underdetermined problem (Vandal et al., 2017; F. Wang et al., 2021). Recent studies have extended the spatial super-resolution approach to the temporal domain and generated a single image with a fourfold higher spatio-temporal resolution applied to rainfall and temperature data (Serifi et al., 2021). CNNs have also shown their potential in downscaling low-resolution climate model outputs while outperforming other statistical approaches (Baño-Medina et al., 2020; Mu et al., 2020; Sun & Tang, 2020; Vaughan et al., 2022).

Recently, conditional generative adversarial networks (cGANs) (Mirza & Osindero, 2014) have been becoming increasingly popular for data generation problems. In comparison to classical CNN approaches, their advantages are that they do not rely on a predefined expert metric, but instead utilize an evolving metric in the form of an individual trained neural network. Furthermore, they have a stochastic design which enables them to generate an ensemble of solutions (Goodfellow et al., 2014). cGANs consist of two networks: a generator and a discriminator. The generator, typically a CNN, generates high-resolution images conditioned on low-resolution inputs, whereas the discriminator evaluates the quality of the generated images by distinguishing between real and artificial images. The generator´s task trying to trick the discriminator is defined by the model´s objective function (Ledig et al., 2017; X. Wang, Yu, et al., 2019). Both networks are simultaneously trained in an adversarial manner. This concept of a two-part architecture and model training has increased the generative performance of neural networks significantly, which is illustrated by the creation of realistic human faces (Karras et al., 2019). In climate science, cGANs can learn to reconstruct high-resolution solutions from climate model outputs and random components. Leinonen et al. (2021) demonstrated the performance and capability of cGANs within a spatial super-resolution approach by downscaling coarsened precipitation data from a resolution of 16 km to 1 km. The same idea has also been applied to downscaling global precipitation forecasts (Price & Rasp, 2022; L. Harris et al., 2022). Furthermore, cGANs outperformed traditional precipitation nowcasting algorithms (Ravuri et al., 2021).

Mapping low- to high-resolution precipitation data is an underdetermined problem due to fluctuations across scales. Resolving the temporal evolution of precipitation events in terms of intensity and advection, is necessary to obtain a complete picture of the high variability of precipitation and the expression of extreme events. Kashinath et al. (2021) refer to the generation of spatially and temporally coherent fields as the holy grail of downscaling. However, existing deep learning methods for spatio-temporal downscaling using CNN based downscaling methods can not sufficiently represent the high variability of precipitation due to their deterministic nature. Even though cGANs have proven to be suitable to present a probabilistic solution for the problem, the focus so far has been on increasing spatial resolutions without temporal downscaling. Often, the super-resolution approaches also address spatial or temporal scales not directly transferable to global climate model data. Furthermore, "recurrent structures" such as reappearing local biases in the generated fields can be an issue. This will also be addressed later in this manuscript.

In this study we propose spateGAN, a cGAN for spatio-temporal downscaling of precipitation based on the video super-resolution approach. We compare a deterministic version of the model to a probabilistic version. Precisely, the objective of this study is:

1. To evaluate the ability of a 3D fully-convolutional cGAN to simultaneously downscale rainfall fields in space and time, from a spatial resolution of 32 km to 2 km and temporally from 1 hr to 10 min.
2. To analyze the model results with respect to spatial structures, temporal consistency and extreme value statistics of the generated fields.

2 Methods

In the following we introduce a new spatio-temporal downscaling approach using a conditional generative adversarial network that learned to downscale spatially and temporally coarsened gridded precipitation observations from a weather radar network (Figure 1). As an evaluation case study we applied the final trained models to the domain of whole Germany and a time period consisting of 12 weeks of data distributed over all seasons. We compared a deterministic and a probabilistic cGAN (spateGAN$_{det}$ and spateGAN$_{prob}$) to a classical CNN approach and trilinear interpolation.

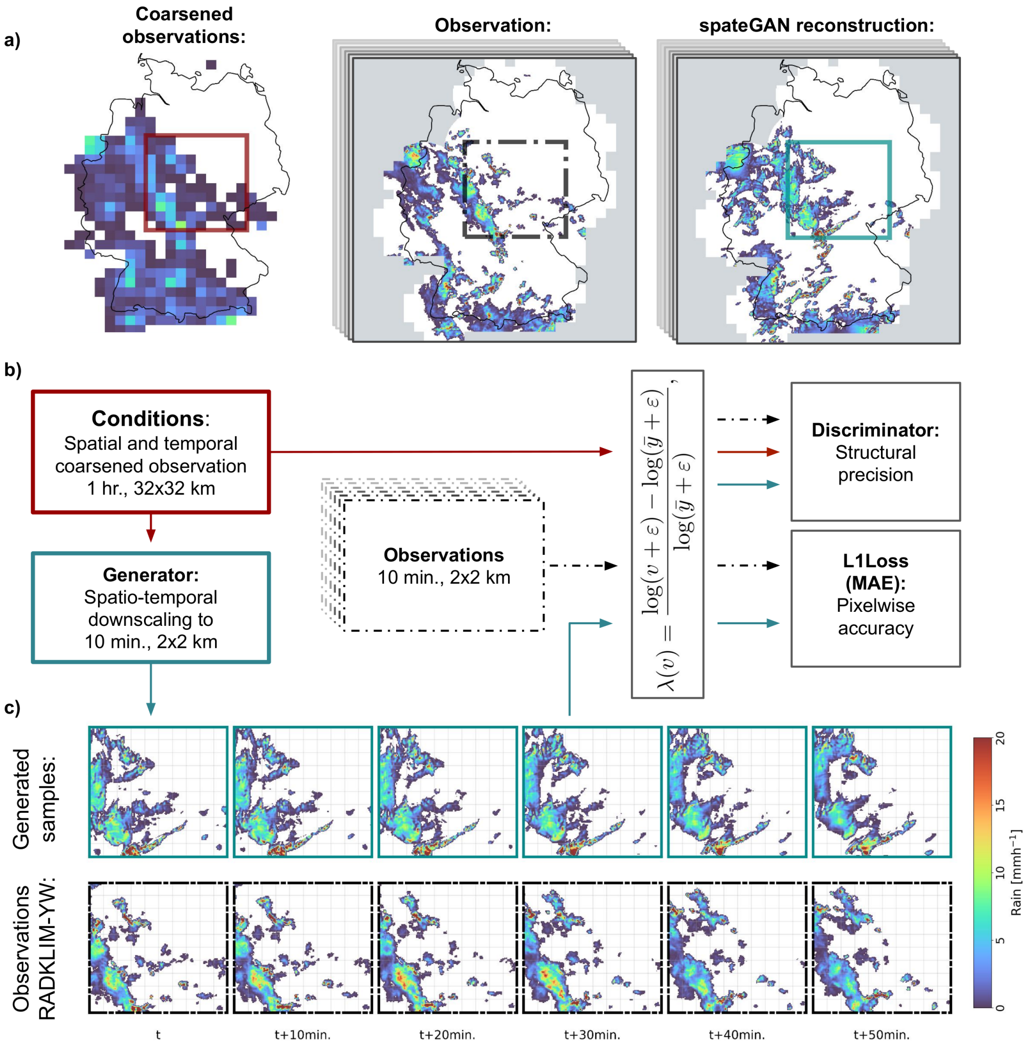

Figure 1. Overview of the proposed spateGAN model for spatio-temporal downscaling of precipitation data. The Figure illustrates the downscaling of a complex precipitation event in Germany, with both stratiform and convective elements. (a) spateGAN downscales coarsened data, derived from weather radar images, with arbitrary spatial and temporal dimensions from a resolution of 32x32 km and 1 hour to a higher resolution of 2x2 km and 10 minutes. The model is trained on smaller patches, represented by the colored boxes. (b) Schematic overview of the model components and training process. (c) Detailed downscaling results from a). spateGAN$_{det}$ is able to convert the hourly resolved coarsened data into a sequence of temporally consistent, finely structured precipitation fields, while also reconstructing the original distribution with higher precipitation intensities.

2.1 Conditional Generative Adversarial Networks for Downscaling

A conditional generative adversarial network comprises two neural networks, the generator G and the discriminator D, which are trained in an adversarial manner. G is a function

$$\begin{array}{rcl} G:\mathbb{R}^{t\times n\times m} & \rightarrow & \mathbb{R}^{d_t t\times d_s n\times d_s m} \\ x & \mapsto & G(x) \end{array} \tag{1}$$

that performs the actual spatio-temporal downscaling of the coarse input x by increasing the temporal resolution by a factor $d_t \in \mathbb{N}$ and the spatial resolution by a factor $d_s \in \mathbb{N}$. In this study $d_t = 6$ and $d_s = 16$. The number of time steps t and grid cells n, m were fixed during training, but can be larger during inference. The discriminator D is a classifier

$$\begin{array}{rcl} D:\mathbb{R}^{t\times n\times m}\times\mathbb{R}^{d_t t\times d_s n\times d_s m} & \rightarrow & \mathbb{R} \\ (x,y) & \mapsto & b \end{array} \tag{2}$$

that distinguishes whether the sequence of high-resolution rainfall maps y has been artificially generated from x (i.e. $y = G(x)$) or is the original high-resolution radar image corresponding to x (Figure 1, b). Both functions are defined as convolutional neural networks (see Section 2.2) trained in a so called adversarial training process. G and D improve their abilities, the generation and discrimination of realistic rainfall time sequences by alternatively minimizing and maximizing the objective function described in Section 2.3. The key point is the custom trainable objective function for G which does not require prior knowledge about the problem to be constructed, but is learned from the data itself via D. The data set and its preparation is explained in Section 2.5. The selection of an optimal model during training and its evaluation requires metrics that we introduce in Section 2.6.

Opposed to the downscaling task is the coarsening operator that was used to synthetically produce coarsened data from high-resolution images. We can define it by

$$\begin{array}{rcl} C:\mathbb{R}^{d_t t\times d_s n\times d_s m} & \rightarrow & \mathbb{R}^{t\times n\times m} \\ y & \mapsto & C(y), \end{array} \tag{3}$$

where $C(y)_{i,j,k} := \frac{1}{d_t d_s^2}\sum\limits_{i'=i}^{i+d_t}\sum\limits_{j'=j}^{j+d_s}\sum\limits_{k'=k}^{k+d_s} y_{i',j',k'}$ is the average over d_t time steps and d_s by d_s grid cells. If not mentioned otherwise we will refer to y as the original high-resolution observation image that was used to produce x, i.e. $x = C(y)$.

2.2 Network Architecture

G and D are convolutional neural networks with a model architecture (Figure 2 a) built from three principal functional blocks (Figure 2 b). G is fully convolutional. The final architecture resulted from an iterative model optimization with special focus on spatio-temporal consistency and the absence of recurrent structures and artifacts. Due to the training time of several days, a full hyperparameter tuning routine and ablation study had to be omitted. For both networks we included 3D convolutional layers. For D these allow the extraction of spatio-temporal features of rain field structures for the decision making. For G they allow to account for spatial and temporal non-linear correlation embedded in the given conditions (Tran et al., 2015) and the reconstruction of temporally consistent high-resolution rainfall fields.

Convolutional-Block

The *Convolutional-Block* is intended to efficiently represent spatio-temporal structures within a feature map. The first part processes the input data through a 3D convolutional layer with kernel size $1\times1\times1$. Depending of the previous layer, the feature dimensionality is decreased to save computational costs and allow for a deeper model (Szegedy et al., 2015). This is followed by a ReLU activation function, another 3D-convolutional layer with kernel size $3\times3\times3$, a Batch Normalization layer and another ReLU activation (Ioffe & Szegedy, 2015).

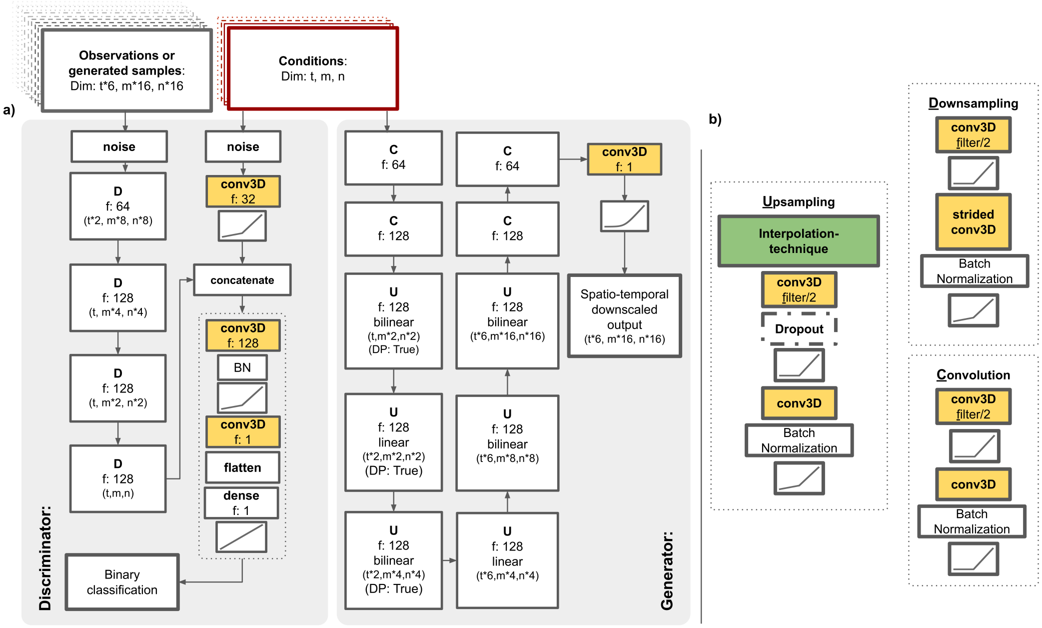

Figure 2. Detailed model architecture of spateGAN consisting of a generator and a discriminator. (a) The discriminator acts as a classification model, evaluating whether the high-resolution time sequences it receives are real or artificial, taking into account their possible affiliation with the coarsened input data provided as a condition. The generator spatially and temporally downscales the coarsened input data. For spateGAN$_{prob}$ dropout layer within the first three *Upsampling-Blocks* enable ensemble generation. (b) Architectures of *Upsampling, Downsampling and Convolutional Blocks*, the main components of both networks.

Upsampling-Block

The upsampling part of the network intends to increase the resolution of the input data by refining the grid size using bilinear interpolation in the spatial dimensions and linear interpolation for the time dimension. Each interpolation step is followed by a *Convolutional-Block* using a leaky ReLU activation to prevent the complete inactivity of these layers.

Downsampling-Block

The *Downsampling-Blocks* are only used within the discriminator. They are based on the presented *Convolutional-Blocks*, but with a kernel size of $4 \times 4 \times 4$ within the second 3D convolutional layer combined with strided convolution and leaky ReLU as second activation function. The approach is similar to Isola et al. (2017) and uses the spatial and temporal stride operation to reduce dimensionality of extracted features.

Generator

The generator initially consists of two *Convolutional-Blocks* without Batch Normalization. Subsequently, the spatial and temporal resolution of the hidden representation is increased using six *Upsampling-Blocks* to achieve the factors $d_t = 6$ and $d_s = 16$ to increase the temporal resolution of 1 hr to 10 min and the spatial resolution from 32 km to 2 km. Each interpolation step is followed by a *Convolutional-Block* to adjust spatio-temporal structures. There are two final *Convolutional-Blocks*, where the second block has no Batch Normalization. The model output is determined by a final convolutional layer to reduce the filter dimension. A softplus activation function limits the distribution of the output to positive values, which can be directly interpreted as rainfall intensity in mm/10 min. For each convolutional layer within G with a kernel size > 1 we applied a reflection padding strategy to reduce boundary errors.

Since downscaling is in general an underdetermined problem, the model uncertainty is closely related to the possible valid realizations of the high-resolution image. The capability of ensemble generation can provide additional valuable information. Leinonen et al. (2021) have shown that for pure spatial downscaling noise, passed as an additional generator feature, is suitable for ensemble generation. We compared a deterministic cGAN approach (spateGAN$_{det}$) to an alternative probabilistic approach (spateGAN$_{prob}$) for ensemble generation, exploiting dropout layers (Isola et al., 2017) within the first three generator *Upsampling-Blocks* during model training and inference. The dropout rate was set to 0.2 with temporal constant selected neurons for each individual ensemble member.

Discriminator

One challenge in training the discriminator is that the given data should be distinguished solely based on the temporal and spatial structures and the distribution. As a first model layer we add noise following a Gaussian distribution (mean=0, stddev=0.05) to the high- and coarse-resolution data to counteract a decision making based on a potential numerical inexactness of the generator while the real images are quantized and a perfect match for the coarse data.

There are two input branches to the network. The high-resolution data is processed by a series of four *Downsampling-Blocks*. The first one has no batch normalization layer. The extracted features are concatenated with the coarsened model input data, that passed through one 3D convolutional layer and a leaky ReLU activation function. After another 3D convolutional layer, Batch Normalization and a leaky ReLU activation function, the filter dimension is reduced using a last 3D convolutional layer. The resulting output is flattend and passed to a single dense layer using a linear activation function allowing for binary classification similar to Ravuri et al. (2021). We observed that Batch Normalization would not be required in all downsampling blocks to get to a similar model performance. However, they lead to a faster desirable model state during training (Ioffe & Szegedy, 2015).

2.3 Objective Function

We express the objective functions for spateGAN following Isola et al. (2017) combining Binary Cross Entropy with a L1 loss term. The L1 loss term or mean absolute error is a pixel-wise error that is only applied to the generator objective. It ensures that the generated rain fields remain close to the ground truth. However, the distribution of rainfall deviates strongly from prominent ANN image data sets. Common methods to achieve a well-performing model and a stable training in spite of this, are data logarithmization and normalization routines (L. Harris et al., 2022; Leinonen et al., 2021; Price & Rasp, 2022).

This, however, can amplify the generation of unrealistically high rainfall intensities in case of a model overestimation during inference or training and a potential necessity of a limitation of the value range in form of an activation function like *sigmoid* or *tanh*, or by a fixed allowed maximum value. In our opinion such a constraint would limit the model to perform well in a non-stationary system. Therefore, we present a new alternative approach using an updated objective function. We logarithmized and normalized data that enter the discriminator or were considered for the calculation of the L1 loss according to

$$\lambda(v) = \frac{\log(v + \varepsilon) - \log(\bar{y} + \varepsilon)}{\log(\bar{y} + \varepsilon)}, \tag{4}$$

where $\bar{y}$ is the maximum of the high-resolution pixel values of the training data set (see Section 2.5.2) and $\varepsilon = 10^{-3}$.

The generator, on the other hand, as visualized in Figure 1 b), was provided unmodified input data and also produced output values that follow the original distribution of the radar data set. The final objective function is

$$\begin{aligned}\mathcal{L}_{cGAN}(G, D) \quad = \quad & \mathbb{E}_{x,y}[logD(\lambda(x), \lambda(y))] + \\ & \mathbb{E}_{x}[log(1 - D(\lambda(x), \lambda(G(x))))] + \\ & \alpha\mathbb{E}_{x,y}[||\lambda(y) - \lambda(G(x))||_1] \end{aligned} \tag{5}$$

where G tries to minimize this objective and the adversarial D tries to maximize it. We set α to 20, to align the loss terms to a comparable range. For spateGAN$_{prob}$ we consulted one random ensemble member per training step during model training for loss calculation to save computational resources.

2.4 Comparison Models: Trilinear Interpolation and Convolutional Neural Network

As a baseline model we refined the grid size of the coarsened validation data correspondingly by a spatial factor of $d_s = 16$ and temporal $d_t = 6$ using trilinear interpolation. In addition, we compared the performance of the spateGANs with a classical neural network approach. For this purpose, we trained a CNN with the exact same architecture as the generator of spateGAN$_{det}$ (see Section 2.2) only applying L1 loss from [5] without D. The remaining training routine was unchanged.

2.5 Radar Data

For model training, testing and validation we used RADKLIM-YW, a publicly available gauge-adjusted and climatologically-corrected weather radar product provided by the German Meteorologic Service (DWD) that can be retrieved from Winterrath et al. (2018) . The radar composite contains information of 16 weather radars adjusted by approx. 1000 rain gauges homogeneously distributed throughout Germany. A detailed description of the radar data processing and correction can be found in Winterrath et al. (2017).

The grid extent is $900\,\text{km} \times 1100\,\text{km}$ with a resolution of $1\,\text{km} \times 1\,\text{km}$. The temporal resolution is 5 minutes, where each grid cell represents a 5 minute rainfall sum. Regions not covered by the 150 km measurement radii of the radars or missing measured values are marked with "NaNs". For our investigation we used data from 1 January 2010 until 31 December 2021. After downloading we transformed the binary data to a NetCDF format following Chwala and Polz (2021) to be able to easily handle the large amounts of data (1Tb/year).

To prevent information leakage and to validate the model's ability to generalize outside the training distribution, the data was split into three sets: 2010–2019 for training, 2020 for testing, and 2021 for validation. All presented results stem from the validation data set.

2.5.1 Data Preprocessing

Before network training, testing and validation, suitable data was selected, the downscaling factor was defined and the high-resolution samples were coarsened. The spatial resolution should increase 16-fold from $32 \times 32\,\text{km}$ to $2 \times 2\,\text{km}$ and the temporal resolution 6-folded from 1 hour to 10 minutes. The chosen scales are sufficient to simulate the downscaling of global climate model data, which can be provided with similar resolution and to be fine enough to reveal the high temporal and spatial variability of precipitation. A further increase of the resolution towards the original RADKLIM-YW data ($1 \times 1\,\text{km}$ and 5 min) would have exceeded our currently available computational resources in terms of GPU memory. Consequently, as a first preprocessing step, the data was spatially averaged and temporal aggregated to a 2 km and 10 minute resolution.

2.5.2 Training and Testing Sample Preparation

GPU memory limitation did not allow the usage of longer time series of whole maps of Germany for model training and testing. Therefore, we randomly selected samples with a spatio-temporal extent of 160×160 pixels and 36 time steps, i.e. $320\,\text{km} \times 320\,\text{km} \times 6\,\text{hr}$. This approach also reduces the risk of the model memorizing spatial dependencies and patterns in the data.

The rain intensity in the data follows a near-lognormal distribution and only about 5 % of the pixels of the radar composite contain precipitation, leading to a high imbalanced and skewed distribution which is difficult for training neural networks. The main issue is learning reasonable predictions for the minority class (J. M. Johnson & Khoshgoftaar, 2019). For rainfall this refers to rarely occurring events and high precipitation intensities. To overcome this problem a simple data augmentation routine was applied. This routine balances the distribution of the train and test samples, increasing the number of wet pixels and total amount of precipitation, and allowing the model to focus on relevant rain events. The data augmentation process selected only samples free of missing values, total precipitation (of all time steps and pixels) exceeding 1000 kg and with at least 100 kg/ 10min per time step for 2/3 of all time steps. To avoid a systematic bias due to the prevailing westerly wind flow influence in Germany, half of the chosen samples were rotated (90° or 270°) or mirrored (vertically or horizontally).

In total, 112,500 samples were randomly drawn for model training (y_{train}) and 1000 samples y_{test} for model testing during training. The test data was also used for model selection (see Section 2.8). As a final preprocessing step, coarsened versions $C(y_{train})$ and $C(y_{test})$ were calculated, resulting in a final model input shape during training ($t \times n \times m$) of 6 time steps and 10×10 pixels.

2.5.3 Validation Data

To validate the model performance, we utilized the fully convolutional architecture of G to downscale entire maps of Germany. This entails a future possible application of downscaling global climate model outputs over a larger domain than the training samples dimension, and the model's ability to generalize for this. To include all seasons and connected temporal sequences, while reducing data volume, we selected the first week of each month of 2021 for validation, resulting in 12,096 validation time steps.

We applied $C(y_{val})$ to derive the coarse validation data, ignoring missing values and setting completely empty coarsened pixels to zero. After model prediction, we masked the downscaled data to exclude pixels with NaN values in y_{val} and areas of coarsened pixels that were not entirely within the radar network coverage, but intersect with it. Additionally we excluded the first and last hour of individually predicted time steps to avoid temporal boundary errors. We applied this procedure to contain all available information in the coarsened data, but derive valid predictions only for those areas where no data is missing. Evaluation metrics were calculated for a cropped area of $370 \times 560\,\text{km}$ (highlighted in Figure 6) to further mitigate boundary effects.

The length of time sequences downscaled by G is mutable and only limited by GPU memory. Using a NVIDIA Tesla V100, G is able to predict 66 time steps of high-resolution maps ($66 \times 480 \times 480$) from 11 coarse precipitation maps ($11 \times 30 \times 30$) in one single processing step, taking 0.1 seconds. Successive predictions were made for contiguous time sequences of this size, resulting in 11,652 images. For spateGAN$_{prob}$ we calculated, according to Section 2.2, 5 ensemble members (spateGAN$_{prob01,02etc.}$) using fixed drop-out neurons for each member and a sixth member, spateGAN$_{prob06}$, in which the selected neurons were randomly changed for every prediction step, i.e. 6 hours. The aggregation of this mixed ensemble member represents the accumulated ensemble mean in this study.

2.6 Metrics

The high temporal and spatial complexity of precipitation makes it difficult to validate the results using a single metric. In addition, different users and decision makers have different requirements on the capabilities of a downscaling model. Thus, the evaluation of the results was carried out with a set of metrics considering different spatial scales and temporal aggregations. Additionally, a qualitative analysis was performed. For calculating the following metrics and for all shown results, we set observed (R_{ref}) and generated (R_{gen}) rain rates below $0.01\,mm\,h^{-1}$ to zero.

2.6.1 Fractions Skill Score

The Fractions Skill Score (FSS) is a spatial verification method to evaluate the performance of precipitation forecasts. It is a measure of the rainfall misplacement error with respect to a given spatial and temporal scale (N. Roberts, 2008; N. M. Roberts & Lean, 2008). A neighborhood of a pixel P contains all grid cells in a r by r square centered at P and T previous and following time steps. Let f_{ref} be the fraction of grid values larger than δ contained in a neighborhood averaged over all possible neighborhoods in an observed image. We define f_{gen} in the same way using the generated image. Then the FSS for δ, r and T is defined by

$$FSS = \frac{\overline{(f_{gen} - f_{ref})^2}}{\overline{f_{gen}^2} + \overline{f_{ref}^2}}, \tag{6}$$

where $\overline{f}$ denotes the average over all images in the data set. For ensemble predictions the fraction is given by the average fraction over all ensemble members. We computed the FSS for various combinations of thresholds δ and scales, r and T.

2.6.2 Radially Averaged Logarithmic Power Spectrum Density

We computed the radially averaged power spectral density ($RAPSD$) and temporal power spectrum density PSD_t to analyze spatial and temporal patterns independent of their location (D. Harris et al., 2001; Sinclair & Pegram, 2005). The $RAPSD$ of a single image was obtained through transforming its 2D power spectrum into a 1D power spectrum by radial averaging, as implemented in PYSTEPS (Pulkkinen et al., 2019). The pixel wise power spectrum along the time dimension is referred to as PSD_t. We calculated the RAPSD for single images ($RAPSD_{10}$), hourly aggregated images ($RAPSD_{60}$) and the accumulation of the entire evaluation data set $RAPSD_{aggr}$.

We compared the power spectrum density of the artificially generated rain fields with the analog measure derived from the observation data. First, we used $RAPSD_{10}$ to evaluate spatial patterns in terms of their frequency and amplitude. Second, we used PSD_t and $RAPSD_{60}$ to quantify the ability to generate temporally consistent fields. And third, we used $RAPSD_{aggr}$ to reveal if models produce recurrent structures (local biases) that sum up over time and are distinct from recurrent local structures in the reference data. An example of such structures is given in Figure 6.

2.6.3 Point Wise and Distribution Error

As a point wise error we computed the mean absolute error (MAE) given by

$$MAE = \overline{|R_{ref} - R_{gen}|}. \tag{7}$$

The continuous ranked probability score ($CRPS$) is a generalization of the mean absolute error and evaluates a probabilistic models predictive distribution against observed values (Gneiting & Raftery, 2007).

The relative $BIAS$ measures the average model error as a percentage of the mean observed rainfall and is given by

$$BIAS = \frac{\overline{R_{gen} - R_{ref}}}{\overline{R_{ref}}} * 100 \tag{8}$$

The Kolmogorov-Smirnov (KS) test measures the maximal distance between the cumulative distribution of observed and generated rainfall. It evaluates the modelled distribution independent of the spatial distribution of values. Because of the skewed distribution of rainfall this maximal distance is most often located at low rainfall intensities which limits conclusions about extreme values.

2.7 Model Training

Each model was trained for three days resulting in about 3×10^5 training steps using mixed precision. The optimization of the spateGANs followed a standard approach by alternating between one gradient descent step for D, followed by one step for G (Goodfellow et al., 2014) and counted as one training step of the spateGAN. We trained on randomly selected samples from the training data set on one Nvidia Tesla V100 GPU limiting batch size to 7. For gradient descent, Adam optimizer was chosen with a learning rate of 1×10^{-4} for G (momentum parameters: $\beta_1 = 0.0, \beta_2 = 0.999$) and 2×10^{-4} for D ($\beta_1 = 0.5, \beta_2 = 0.999$). Models were saved after every 500th training step to later select the best performing state. We implemented the ANNs and model optimization in a Python framework using TENSORFLOW (version: 2.6) (Developers, 2022).

2.8 Model Selection

We selected the best performing models (i.e. the optimal state of either CNN, spateGAN$_{det}$ and spateGAN$_{prob}$ during training) by downscaling the test data. We took the structural error of all generated images into account using both $RAPSD_{aggr}$ and the average $RAPSD_{10}$. We represent the RAPSD deviation by a single value by calculating the mean absolute error of the logarithmized RAPSDs of predicted and real images:

$$\sigma = \frac{1}{n}\sum_{i=1}^{n} |10 * log_{10}(RAPSD_{real}) - 10 * log_{10}(RAPSD_{predicted})| \tag{9}$$

Based on $RAPSD_{aggr}$, σ_{aggr} considers potential model artefacts in the form of recurrent structures and the model ability to reconstruct adequate rain sums for a longer time period. Based on $RAPSD_{10}$, σ_{10min} takes the models ability to generate rain fields with spatial structures of the right amplitudes and frequencies into account. To avoid too strong influence of boundary errors in this selection we excluded the outermost edge, corresponding to one coarse resolution pixel, for this calculation. Finally, the model minimizing $\sigma_{aggr} + \sigma_{10min}$ was selected.

3 Results

To evaluate the spatio-temporal downscaling performance we considered the models capability to reconstruct the target distribution from spatially and temporally coarsened input data and to generate rain fields that closely resemble the observations regarding spatial structure and temporal consistency.

3.1 Qualitative Analysis

We start with a qualitative analysis examining a detailed visualization of the sequences generated for three rain events. One is a convective case study scenario and the other two show a stratiform and a mixed type rain event. The observation data, their

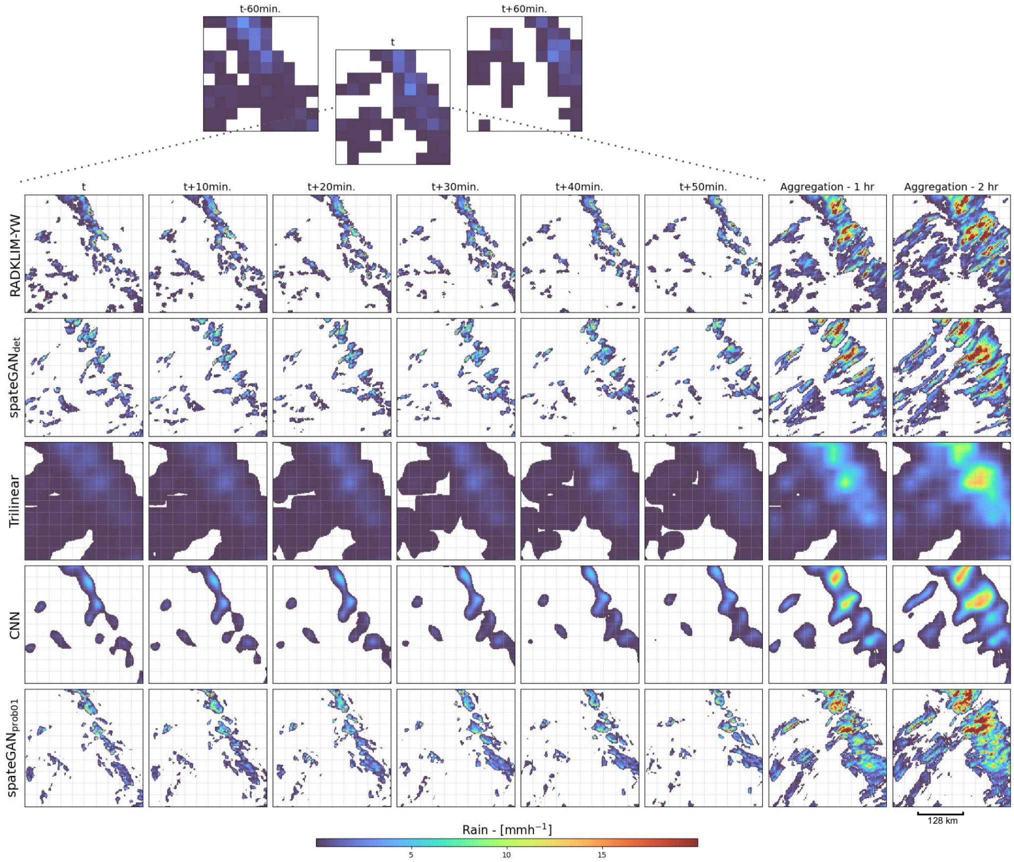

Figure 3. Detailed case study of the spatio-temporal downscaling performance for a convective precipitation event for central Germany. Shown are a temporal sequence of coarsened model input data, associated RADKLIM-YW observations, and model predictions. Hourly and two-hourly aggregated images highlight specific advection structures.

associated coarsened representation and the respective models are shown in Figures 3, 4 and A1. The predictions from the probabilistic generative approach stem from a single ensemble member (spateGAN$_{prob01}$). Additionally, the preceding and subsequent time steps of the coarsened images are presented to provide a better understanding of what information is available to the model to generate the high-resolution images. A more complete picture is given by the attached animations visualizing the full time sequences of different events (`https://doi.org/10.5281/zenodo.7636929`).

Case Study: Convective Rain Events

Figure 3 shows the temporal evolution of a convective rainfall event. The challenge for the downscaling models was to determine that the connected rainfall field in the coarsened input data represents disconnected convective cells and to localize them correctly with plausible advection.

Both spateGAN approaches effectively generated small convective rain cells from the low-resolution data which cannot be easily identified as artificially generated. The spatial structures, localization and advection were in good agreement with the observation data. However, there are differences in certain regions. For example, a more connected rain field in the north was represented as smaller separated cells. The observed small rain event in the southeast at t+20 min with a rain rate $> 20\, mm\, h^{-1}$ was generated as a larger event with lower rain rates. Despite these small scale dissimilarities, spateGAN was able to construct plausible local extremes like in the northern part of the images. In addition to the individual time steps, the 1-hour aggregations revealed advection structures that are very similar to the observation data in large parts of the images. This supports the hypothesis that the model is able to reproduce spatio-temporally consistent small-scale rainfall structures with plausible advection.

The CNN could generate rain fields with reasonable position and timing, but the cells lacked fine-scaled spatial structure and local extremes. Especially the gradients were very smooth. The model was not able to separate individual convective cells, however by comparing the presented time steps in chronological order, a plausible movement and temporal consistency became apparent.

The trilinear interpolation created a blurry version of the low-resolution data lacking local gradients, extreme values or advection.

Case Study: Stratiform Rain Events and Embedded Convection

Figure 4 presents the one hour time sequence of a stratiform rain event. The challenge for the models was to reconstruct the evolution of this larger rain field including areas with no precipitation and a smaller separated cell in the north, from contiguous pixels in the coarsened input data. The results from the spateGANs appear very similar to the observational data, including the size and positioning of the generated rain fields. The artificially generated events show plausible structures with a slight underestimation of the maximum rainfall intensity in, e.g., image t+20min. Higher rainfall intensities in the southeast corner and correctly positioned holes were created. The small detached rain events in the north are also depicted and are hardly distinguishable from the observation data. The generated structures exhibit a plausible temporal and spatial development, even though the rain field is moving slowly. spateGANs ability to generate both small and large rain events in a single image is further demonstrated for a complex precipitation event in Figure A1.

As within Figure 3, the trilinear interpolation and CNN results were blurry and lacked spatial structure. The CNN was more accurate in terms of the spatial extent of the rain field, while the trilinear interpolation produced fields that exceeded the spatial extent of the reference.

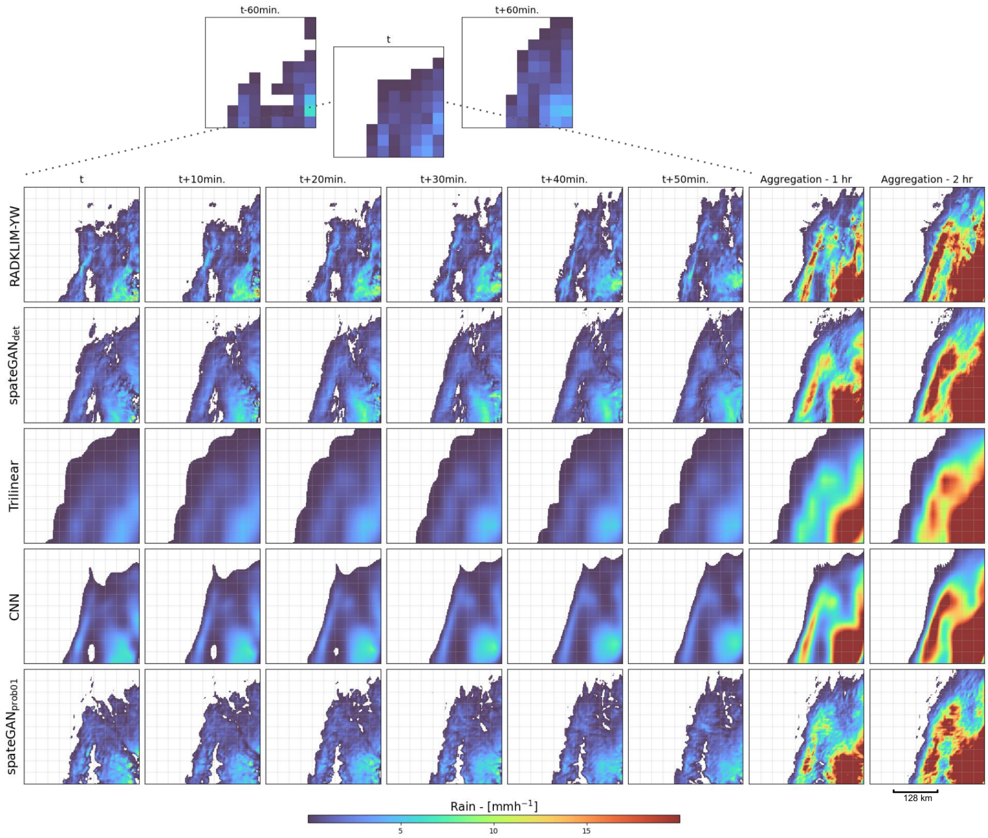

Figure 4. As Figure 3 for a stratiform event.

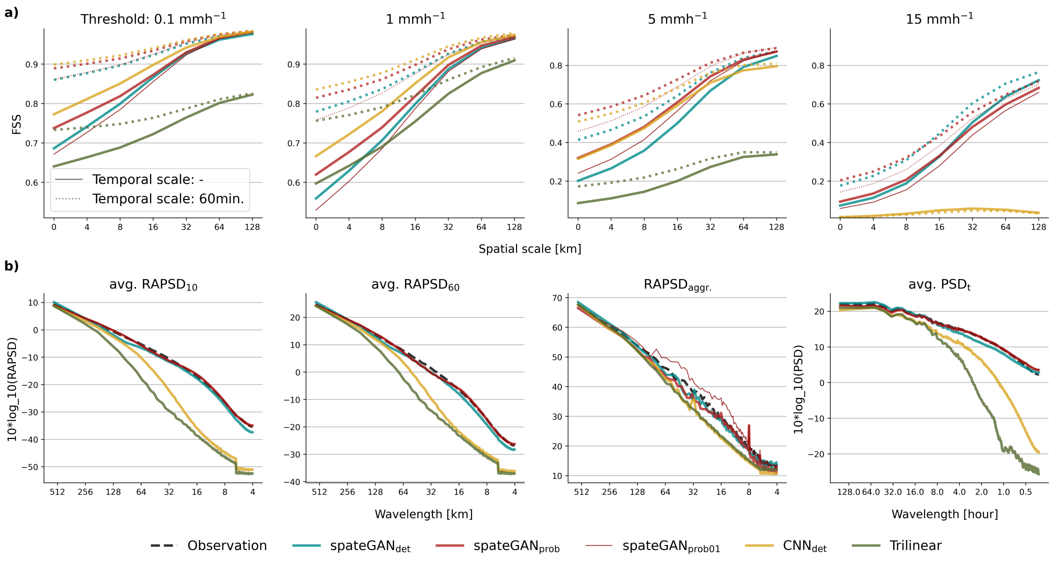

Figure 5. Evaluation of the downscaling methods (spateGANs, CNN, and trilinear interpolation) for a cropped area of the 2021 validation data set for Germany. (a) presents the Fractions Skill Score (FSS) for different thresholds and spatial and temporal scales, with the ensemble FSS of multiple members for spateGAN$_{prob}$. Part (b) evaluates the generated spatial and temporal structures using power spectra analysis. spateGAN$_{prob}$ refers not to multiple ensemble members, but to the mixed ensemble member as described in Section 2.5.3. The temporal consistency of the generated fields is evaluated using $RAPSD_{60}$ and the average PSD_t. All ANN models show peaks in $RAPSD_{aggr.}$ at different wavelengths and intensities, indicating the presence of recurrent patterns in the predictions.

3.2 Quantitative Investigation

The quantitative analysis is divided into two parts. First we investigated the models regarding their capability to generate detailed spatio-temporal rain field structures by analyzing the power spectrum. Then, we examined the pixel accuracy and the ability to reconstruct a skillful distribution in time and space by calculating the FSS, CRPS, MAE, KS statistics and BIAS.

3.2.1 Structural Analysis

We calculated the average $RAPSD_{10}$ and $RAPSD_{60}$ of the high-resolution observation images and the associated model predictions to investigate whether the models are able to represent the structural variability and advection of precipitation across spatial and temporal scales. The same analysis was performed for the accumulated precipitation of all 11652 validation images ($RAPSD_{aggr.}$) to visualize potential undesirable model characteristics such as the generation of recurrent structures that would manifest as peaks at certain wavelengths.

Figure 5 b) shows that the generated images from spateGAN$_{det}$ and spateGAN$_{prob}$ have a high structural similarity to the observations for both, single images and hourly aggregations on all considered scales. A small underestimation occurred between wavelengths of 128 to 64 and < 6 km for spateGAN$_{det}$. Respectively a slight overestimation occurred for spateGAN$_{prob}$. The same was observable in the temporal power spectrum PSD_t for wavelengths between 30 min. and 4 hours. For higher frequencies spateGAN$_{prob}$ showed a slight overestimation. The $RAPSD_{aggr}$ was close to the observation data. How-

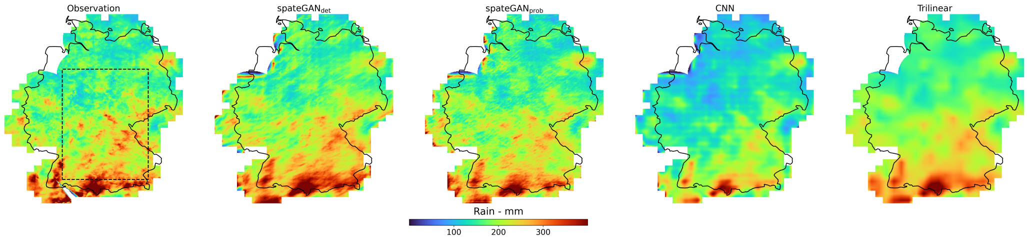

Figure 6. Aggregated observed and predicted rainfall of the validation data set for Germany for the year 2021. The accumulation shows the models ability to maintain the total rainfall amount and reveals recurrent structures within the predictions that contradict the physical principle of developing rain fields. spateGAN$_{prob}$ represents an ensemble mean as described in Section 2.5.3 and the rectangle defines the area considered for the quantitative analysis.

ever, peaks mainly prominent at a wavelength of 8 and 6 km could be observed. Recurrent structures with this frequency were also visible in the accumulated rainfall maps from Germany in Figure 6. Predictions of spateGAN$_{det}$ also exhibited this conspicuity at a wavelength of 32 km. At shorter aggregations (e.g. individual predictions, $RAPSD_{10}$ or $RAPSD_{60}$) these structures were not detectable.

For the CNN, $RAPSD_{10}$, $_{60}$ and $_{aggr.}$ showed an underestimation, especially for higher frequencies. This results from the missing model ability to generate small scale structures and to reconstruct the original high-resolution distribution. Recurrent structures could be also observed at wavelength of 32 km.

Trilinear interpolation was in general not capable to generate small scale spatio-temporal structures that were similar to the observation data. A high RAPSD and PSD_t underestimation could be shown for wavelength smaller 128 km or 8 hours. Within the whole accumulated validation data set no recurrent structures could be observed considering $RAPSD_{aggr}$ or Figure 6.

3.2.2 Distribution Reconstruction Skill

The coarse resolution provided as model input compresses the distribution of rainfall intensities towards lower values. The decisive factor of a skilful downscaling model is therefore not only the generation of realistic spatial structures, but rather the ability to reconstruct the correct distribution of rainfall intensities with accurate spatial and temporal placement of the rain events. We measured this downscaling skill by considering the FFS for the spatial and temporal precision of reconstructing high intensities using thresholds δ of $0.1\,mm\,h^{-1}$, $1\,mm\,h^{-1}$, $5\,mm\,h^{-1}$ and $15\,mm\,h^{-1}$. These thresholds represent the 0.9, 0.97, 0.997 and 0.9998 quantiles of the validation data set. The spatial scales r were between 0 and 128 km and the temporal scales T were 0 and 60 minutes. The results are shown in Figure 5 a). The generative models demonstrated a high skill for small to moderate rainfall (0.1 and $1\,mm\,h^{-1}$) with FSS exceeding 0.9 at a spatial scale of 32 km. They also performed well for high and strong rainfall intensities, with FSS values over 0.8 and 0.7 for a threshold of 5 and $15\,mm\,h^{-1}$. The score of spateGAN$_{prob}$ increased further, especially for small rain rates and scales, when multiple ensemble members were considered and the ensemble FSS was calculated. The CNN showed the best performance for small and moderate rainfall rates, but the accuracy decreased for strong rainfall intensities with a maximum FSS of 0.06 for $15\,mm\,h^{-1}$. Trilinear interpolation performed well for moderate precipitation ($1\,mm\,h^{-1}$) but had the lowest overall skill.

Additionally, we calculated pixel accuracy metrics CRPS, or MAE for deterministic models, and the BIAS, as well as the distribution error as the KS statistics shown in Table 1. In terms of MAE, KS statistics, and BIAS the spateGAN models achieved

Table 1. Set of downscaling skill metrics computed for the validation data set. The FSS refers to the maximum score of Figure 5 a) each model achieved for different thresholds. For spateGAN$_{prob}$ multiple ensembles were considered for CRPS and FSS, a single member for MAE, KS statistic, power spectra deviation σ_{10min} [9] and BIAS.

	CRPS/MAE	KS	FSS$_{0.1}$	FSS$_1$	FSS$_5$	FSS$_{15}$	σ_{10min}	BIAS
spateGAN$_{det}$:	-/0.018	0.010	**0.98**	0.97	0.87	**0.73**	1.36	3.35
spateGAN$_{prob}$:	**0.012**/0.018	0.014	**0.98**	0.97	**0.89**	0.71	**0.31**	-3.55
CNN:	-/**0.012**	**0.008**	**0.98**	**0.98**	0.81	0.06	16.1	-22.22
Trilinear:	-/0.016	0.20	0.81	0.91	0.23	0	18.6	**-0.25**

overall good scores, compared to CNN and trilinear interpolation. The BIAS of spateGAN$_{det}$ showed a slight overestimation and an underestimation for spateGAN$_{prob}$. The CNN had the best KS score and MAE, but a negative BIAS of -22.28 % indicated a strong underestimation (see Figure 6). Trilinear interpolation showed the best BIAS with -0.28 % .

3.3 Ensemble Downscaling

The generation of multiple ensemble members is crucial to quantify uncertainties in the downscaling process like the likelihood of extreme events (Pathak et al., 2022).

By comparing the probabilistic generative approach to the deterministic, it could be shown that the predictions of an individual ensemble member, like spateGAN$_{prob01}$, looked similarly realistic as the predictions of spateGAN$_{det}$ (see Figure 3, 4 and A1). Regarding the RAPSD$_{10}$, RAPSD$_{60}$ and PSD$_t$ the predictions where even closer to the observation data as can be seen in Figure 5. The downscaling skill of spateGAN$_{prob01}$ was only minimally reduced with lower FSS for the thresholds 0.1, 1 and 15 $mm\,h^{-1}$, but higher scores for 5 $mm\,h^{-1}$. The potential of a probabilistic approach which considers multiple spateGAN$_{prob}$ ensemble members was investigated by calculating the CRPS and ensemble FSS (see Table 1). The CRPS showed an improvement with a value of 0.012 compared to the MAE of SpateGANdet and SpateGAN$prob$01. Furthermore, the FSS indicated a better downscaling performance compared to SpateGANdet and SpateGAN$prob$01, particularly for small scales and low rainfall amounts. The probabilistic model was also able to well represent the precipitation sum of the validation reference considering the aggregated ensemble mean, as can be seen in Figure 6.

However, Figure 5 shows that the aggregation of a single ensemble member (RAPSD$_{aggr}$ for spateGAN$_{prob01}$) showed an overestimation from scales between 8 and 128 km. We assume that this model characteristic was due to the chosen dropout routine. For one ensemble member selected drop out neurons were fixed for all time steps. The behaviour was not visible in single predictions and could only be revealed via the aggregation and analysis of multiple thousand images. To address this constraint, we emphasize to always consider multiple ensemble members, when applying this approach for longer time series.

Furthermore, we experimented to change the drop out rate after model training, which lead to an increased variance of the ensemble members. However, the downscaling skill was not further improved. Additionally, we trained a model applying random drop out neurons for each time step, which could generate temporal consistent rain fields without issues when aggregating single ensemble members. However, it frequently produced low rain rates during dry time steps and regions. Overall this exemplifies that various approaches for ensemble generation are feasible, but the creation of ensembles that reflect the physical plausible solutions and the stochasticity of the target data set is challenging and therefore subject to further research.

4 Discussion

In this study we proposed spateGAN, a novel approach for spatio-temporal downscaling of precipitation data combining cGANs, 3D convolution and interpolation techniques. It effectively increases the spatial resolution of coarsened weather radar data from 32 km x 32 km and 1 hour to 2 km x 2 km and 10 minutes. In the following we will discuss the models ability to accurately reconstruct spatial structures with temporal consistency and correct extreme value statistics. Additionally, we present the models limitations and additional unexpected findings.

Spatial Structures

The qualitative investigation (see Section 3.1) and the presented animation prove the ability of spateGAN to generate plausible precipitation fields from coarsened input data that are hardly classifiable as artificially generated. This is supported by the power spectrum analysis using $RAPSD$ and PSD, which are in highest agreement with the observation data for all scales when compared to CNN and interpolation. The FSS confirms that unlike trilinear interpolation and a classical CNN approach, the cGAN approach accurately produces structures with higher rainfall intensities. spateGAN is the only model that is able to generate rain cells of small spatial extent (see Figure 3). Besides the spatial extent and the rainfall intensity, the number of generated cells has a similar order of magnitude compared to the observations. Only the precise location of these cells deviates due to the stochastic nature of the model. spateGAN also tends to produce slightly smoother structures than the observed ones for large scale rain events like shown in Figure 4. We assume that an increase of the training sample dimensions could improve the structural quality of such large rain events. Overall, the results emphasize the necessity of a generative network downscaling approach for modeling realistic rain fields, since trilinear interpolation and CNN lack higher frequencies in the power spectrum. Trilinear interpolation approximates the low-resolution data providing limited additional information, while the CNN generates more detailed, but still too blurry events (Larsen et al., 2016).

Temporal Consistency

The animations of downscaled rain fields illustrate temporal consistency as a key property of spateGAN. The generated fields exhibit plausible advection, showing that rain cells are not randomly appearing and disappearing between time steps. This is supported by the 1 hour and 2 hour aggregations (see case study Figures 3, 4, A1), where the sum of individual time steps leads to smooth, connected cells elongated in the direction of advection. Furthermore, $RAPSD_{60}$ and PSD_t are in high agreement with the observation data. The visual evaluation of the CNN predictions and its improved PSD_t compared to trilinear interpolation also indicate the CNN's ability to generate temporally consistent events. This leads us to conclude that 3D convolutions are suitable for creating temporally coherent downscaled images (Vondrick et al., 2016; Tran et al., 2015). In combination with linear temporal interpolation within G, 3D convolutions are a crucial factor for the generation of these consistently evolving rain fields. 3D convolutional layers in D may also contribute to spateGANs high temporal consistency, which is supported by a similar application for precipitation nowcasting (Ravuri et al., 2021). However, in our use case their impact on structural precision, that is, the localization of rain cells, might be more significant.

Model Limitations

Despite its potential, 3D convolution has certain limitations and its usefulness for video generation is still a matter of debate (Saito et al., 2017). The main challenge is that the possible amount of exploitable large-scale and long-term spatio-temporal correlations is not arbitrarily expandable. It depends on the model architecture and model depth which define the receptive field size. Furthermore, also the spatial and temporal dimensions of the training samples are important, since model extrapolation capabilities beyond this dimension might be highly limited. Overall, the potential is therefore tied to the available GPU resources, while the memory requirements of 3D convolution are substantial. On the other hand, fully convolutional networks allow for arbitrary input dimensions and we found that spateGANs architecture and depth is sufficient to achieve high performance within the super-resolution downscaling approach. While the model predictions are already spatially and temporally consistent beyond the training sample dimensions it remains unclear if the performance could be further increased by leveraging longer time scales and a larger spatial extent during training. We assume that in the case of downscaling global climate data, an increase in the model's receptive field might be crucial to realize the full potential.

Distribution of Downscaled Rainfall

A main objective of a spatio-temporal downscaling model is the ability to accurately reconstruct the distribution of rainfall at a higher spatial and temporal resolution, which is typically characterized by increased variability and extremes. As expected, the FSS of all models declines towards heavier rainfall, which is harder to model due to its rare occurrence and higher spatio-temporal gradients.

Among the evaluated models, spateGAN stands out as the only model that successfully reconstructed rainfall intensities greater than $5\,mm\,h^{-1}$ or $15\,mm\,h^{-1}$, while maintaining a low BIAS ($< 3.6\,\%$). This is a crucial feature that is not provided by the comparison models. Trilinear interpolation shows the lowest BIAS, however, also the lowest downscaling skill in terms of FSS and RAPSD. The CNN predictions show high skill regarding pixel accuracy metrics, distribution error or downscaling skill for small and moderate rain rates. However, the model is not able to skilfully reconstruct strong precipitation intensities. Furthermore, the model fails to preserve the overall rain sum, maintained within the coarsened input data showing a strong negative BIAS (-22.22 %). We therefore emphasize, as also described in Leinonen et al. (2021), that MAE and KS statistics should be interpreted with caution, as the results could be highly affected by the large amount of small values within the skewed rainfall distribution. They are therefore not suitable to account for the model's ability to recover the target rain distribution containing also extreme values. Furthermore, they can lead to poor metrics, even if models are able to generate rain cells with correct structure and intensity, since these rain cells might be slightly off positioned within the underdetermined downscaling problem and the stochasticity of the solution.

Unexpected Findings

Our analysis of long aggregations (several thousand time steps) of generated rain fields revealed the presence of local biases in the form of recurrent structures. With varying intensity and frequency, they could be observed within the predictions of all ANN models. It is known that GANs can produce artefacts (Karras et al., 2019). However, in our case they were not detectable in single images, e.g., by calculating the power spectrum density. Preliminary results indicate that such model behavior is not unique to the models used in this study, as other prominent ANN downscaling models might also be affected by this behaviour.

While the training images for our models are selected at random locations, reducing the influence of topography, the generated structures are not completely random. Instead, they might follow a spatial or even geometric regularity which is contradictory to the physical principle of emerging rain fields. This does not imply that the downscaling performance of the models is reduced, but can be a seen as a limitation and should be a known feature to be tested. In an effort to minimize the occurrence of these structures, we presented a model with a sophisticated architecture and interpolation technique. Furthermore, we also considered the appearance of these structures in the selection process of the final models (see Section 2.8). Despite this, we were unable to completely eliminate them. Our analysis revealed that a discriminator with many parameters (e.g. G : 2 million., D: 10 million) might lead to an earlier and more intense occurrence of these phenomena. Additionally, we assume that the combination of up and down-sampling layers and their kernel sizes also have an influence. To fully understand the underlying mechanisms responsible for the observed structures, a comprehensive investigation involving the comparison of various hyper-parameterizations would be required. Given the computational effort for training one model, this investigation is beyond the scope of this study and will be left for future research. In the geosciences not only single instances, but also the aggregation of many instances is of importance. Therefore, we emphasize that it is not sufficient to only analyze single predictions, but also the models abilities to fulfill global properties like the climatology of the modeled target variable.

5 Conclusion

Downscaling the output of global climate models is a long-standing problem for providing high-resolution information which is needed to develop adaptation and mitigation strategies in a changing climate. We presented spateGAN, a deep generative model, for simultaneous spatio-temporal downscaling of low-resolution precipitation data. The model was trained using ten years of high-resolution country-wide weather radar rainfall observations in Germany. Our results demonstrated that 3D convolution in combination with conditional generative adversarial networks is an effective tool for leveraging spatio-temporal structures embedded in the low-resolution domain to generate temporally consistent high-resolution rainfall fields and reconstruct the scale dependent extreme value distribution with high skill. This confirms that super-resolution deep learning approaches can be extended to the time dimension to map, in addition to the spatial variability, also the temporal evolution of atmospheric variables.

While a visual inspection leads to the conclusion that generated rain cells look realistic, we found the power spectrum analysis and the Fractions Skill Score to be useful metrics for quantifying this property. Pixel accuracy metrics like the mean absolute error were unable to distinguish between models with high or low skill in generating realistic rain fields. Especially our findings about recurrent structures in downscaled rainfall fields show that a structural analysis is very important in order to mitigate these issues. Overall, the chosen analysis was able to prove that models like spateGAN show great potential to complement and even outperform the capabilities of traditional downscaling methods due to their high performance, computational efficiency and the ability to process arbitrary spatial and temporal input dimensions.

One of the primary purposes of spateGAN is the application for downscaling global climate model outputs. We envision that the approach for this task will have to extend the presented video super-resolution approach, since model outputs are biased with respect to the observed precipitation. Therefore, requirements for the downscaling model would include an additional bias correction step. The potential for bias correction and spatial downscaling of weather forecast data using generative networks has in been demonstrated in L. Harris et al. (2022) and Price and Rasp (2022) and resulted in a performance reduction compared to downscaling coarsened observations. A similar result should be expected for spatio-temporal downscaling. However, we assume that with increased lead time a decoupling of model projections from real observations is the reason for the performance decline and not the insufficient potential of the deep learning approach. Additionally, further studies will have to prove if the generated precipitation fields are suitable, e.g. for simulating the characteristics of flood events under future climate conditions. This work should provide a solid basis for such future studies by not only presenting a high performance downscaling model, but also the analytical framework for a comprehensive analysis of the model performance.

Appendix A Supplementary Figure

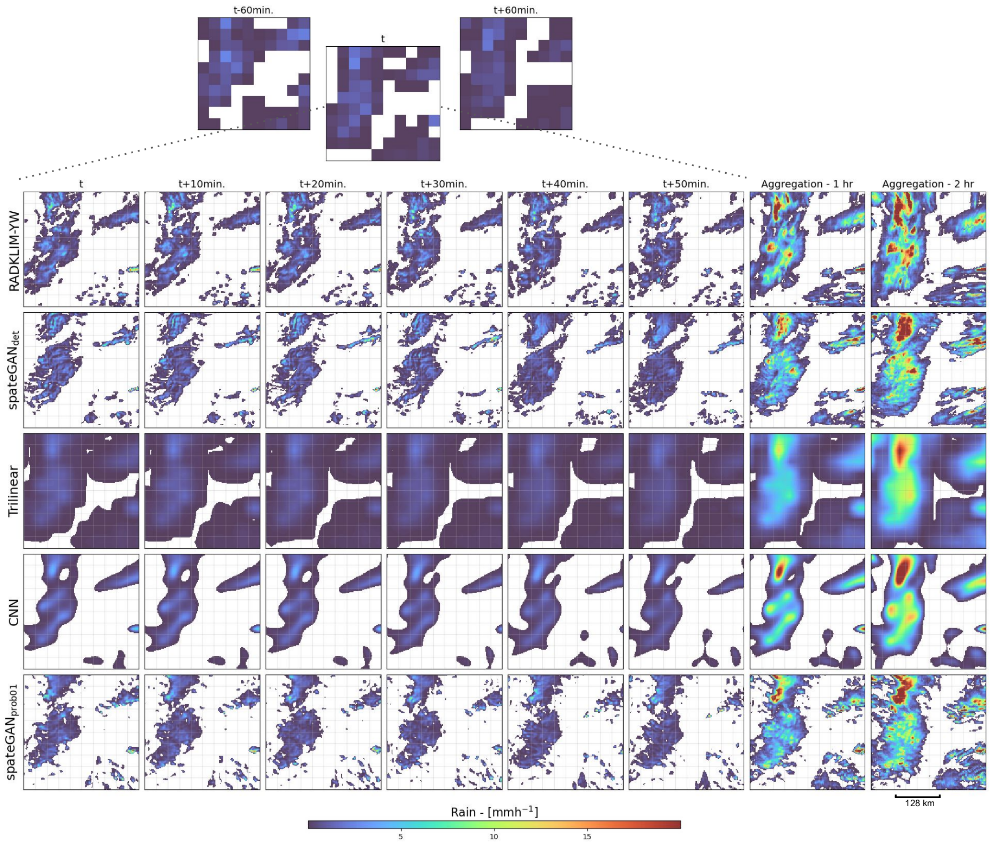

Figure A1. Detailed case study as in Figure 3 for a third event, with a mixture of convective and stratiform rain.

Open Research

The results and models can be reproduced by the publicly available RADKLIM-YW weather radar composite (Winterrath et al., 2018). The CNN and spateGANs were implemented and optimized in a Python framework using TENSORFLOW (version: 2.6) (Developers, 2022). The data and spateGAN models, available in `https://doi.org/10.5281/zenodo.7636929`, provide further insight into the presented spatio-temporal downscaling approach.

Acknowledgments

This study was supported by SCENIC (Storyline Scenarios of Extreme Weather, Climate, and Environmental Events along with their Impacts in a Warmer World) funded by HGF-Innopool, the German Research Foundation (Grant CH 1785/1-2) and the Federal Ministry of Education and Research (Grant 13N14826). We acknowledge support by the KIT-Publication Fund of the Karlsruhe Institute of Technology. Open Access funding enabled and organized by Projekt DEAL.

References

Allan, R. P., & Soden, B. J. (2008). Atmospheric Warming and the Amplification of Precipitation Extremes. *Science*, *321*(5895), 1481–1484. Retrieved 2022-12-20, from https://www.science.org/doi/10.1126/science.1160787 (Publisher: American Association for the Advancement of Science) doi: 10.1126/science.1160787

Baño-Medina, J., Manzanas, R., & Gutiérrez, J. M. (2020). Configuration and intercomparison of deep learning neural models for statistical downscaling. *Geoscientific Model Development*, *13*(4), 2109–2124. Retrieved 2022-09-29, from https://gmd.copernicus.org/articles/13/2109/2020/ (Publisher: Copernicus GmbH) doi: 10.5194/gmd-13-2109-2020

Berg, P., Moseley, C., & Haerter, J. O. (2013). Strong increase in convective precipitation in response to higher temperatures. *Nature Geoscience*, *6*(3), 181–185. Retrieved 2022-12-20, from https://www.nature.com/articles/ngeo1731 (Number: 3 Publisher: Nature Publishing Group) doi: 10.1038/ngeo1731

Chen, D., Rojas, M., Samset, B., Cobb, K., Diongue Niang, A., Edwards, P., ... Tréguier, A.-M. (2021). Framing, Context, and Methods. In V. Masson-Delmotte et al. (Eds.), *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 147–286). Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press. (Type: Book Section) doi: 10.1017/9781009157896.003

Chwala, C., & Polz, J. (2021). *cchwala/radolan_to_netcdf.* Zenodo. Retrieved 2022-12-21, from https://zenodo.org/record/4452204 doi: 10.5281/ZENODO.4452204

Developers, T. (2022). *TensorFlow.* Zenodo. Retrieved 2023-01-18, from https://zenodo.org/record/4724125 doi: 10.5281/ZENODO.4724125

Dong, C., Loy, C. C., He, K., & Tang, X. (2016). Image Super-Resolution Using Deep Convolutional Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(2), 295–307. (Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence) doi: 10.1109/TPAMI.2015.2439281

Gneiting, T., & Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. Retrieved 2022-12-01, from http://www.tandfonline.com/doi/abs/10.1198/016214506000001437 doi: 10.1198/016214506000001437

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative Adversarial Nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (Vol. 27). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

Harris, D., Foufoula-Georgiou, E., Droegemeier, K. K., & Levit, J. J. (2001). Multiscale Statistical Properties of a High-Resolution Precipitation Forecast. *Journal of Hydrometeorology*, *2*(4), 406–418. Retrieved 2022-12-01, from https://journals.ametsoc.org/view/journals/hydr/2/4/1525-7541_2001_002_0406_mspoah_2_0_co_2.xml (Publisher: American Meteorological Society Section: Journal of Hydrometeorology) doi: 10.1175/1525-7541(2001)002⟨0406:MSPOAH⟩2.0.CO;2

Harris, L., McRae, A. T. T., Chantry, M., Dueben, P. D., & Palmer, T. N. (2022). A Generative Deep Learning Approach to Stochastic Downscaling of Precipitation Forecasts. *Journal of Advances in Modeling Earth Systems*, *14*(10), e2022MS003120. Retrieved 2023-02-10, from https://onlinelibrary.wiley.com/doi/abs/10.1029/2022MS003120 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2022MS003120) doi:

10.1029/2022MS003120
IFRC. (2021). *World Disasters Report 2020 \textbar IFRC.* Retrieved 2022-12-17, from https://www.ifrc.org/document/world-disasters-report-2020
Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning* (pp. 448–456). PMLR. Retrieved 2023-02-14, from https://proceedings.mlr.press/v37/ioffe15.html (ISSN: 1938-7228)
Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-Image Translation with Conditional Adversarial Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5967–5976). (ISSN: 1063-6919) doi: 10.1109/CVPR.2017.632
Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer Vision – ECCV 2016* (pp. 694–711). Cham: Springer International Publishing. doi: 10.1007/978-3-319-46475-6_43
Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*(1), 27. Retrieved 2022-12-22, from https://doi.org/10.1186/s40537-019-0192-5 doi: 10.1186/s40537-019-0192-5
Karras, T., Laine, S., & Aila, T. (2019). A Style-Based Generator Architecture for Generative Adversarial Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4396–4405). (ISSN: 2575-7075) doi: 10.1109/CVPR.2019.00453
Kashinath, K., Mustafa, M., Albert, A., Wu, J.-L., Jiang, C., Esmaeilzadeh, S., … Prabhat, n. (2021). Physics-informed machine learning: case studies for weather and climate modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *379*(2194), 20200093. Retrieved 2022-02-02, from https://royalsocietypublishing.org/doi/10.1098/rsta.2020.0093 (Publisher: Royal Society) doi: 10.1098/rsta.2020.0093
Kim, J., Lee, J. K., & Lee, K. M. (2016). Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1646–1654). (ISSN: 1063-6919) doi: 10.1109/CVPR.2016.182
Larsen, A. B. L., Sønderby, S. K., Larochelle, H., & Winther, O. (2016). Autoencoding beyond pixels using a learned similarity metric. In *Proceedings of The 33rd International Conference on Machine Learning* (pp. 1558–1566). PMLR. Retrieved 2023-02-14, from https://proceedings.mlr.press/v48/larsen16.html (ISSN: 1938-7228)
Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., … Shi, W. (2017). Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 105–114). Honolulu, HI: IEEE. Retrieved 2023-02-14, from http://ieeexplore.ieee.org/document/8099502/ doi: 10.1109/CVPR.2017.19
Leinonen, J., Nerini, D., & Berne, A. (2021). Stochastic Super-Resolution for Downscaling Time-Evolving Atmospheric Fields with a Generative Adversarial Network. *IEEE Transactions on Geoscience and Remote Sensing*, *59*(9), 7211–7223. Retrieved 2022-02-23, from http://arxiv.org/abs/2005.10374 (arXiv: 2005.10374) doi: 10.1109/TGRS.2020.3032790
Lucas, A., Katsaggelos, A. K., Lopez-Tapuia, S., & Molina, R. (2018). Generative Adversarial Networks and Perceptual Losses for Video Super-Resolution. In *2018 25th IEEE International Conference on Image Processing (ICIP)* (pp. 51–55). Athens: IEEE. Retrieved 2023-02-13, from https://

ieeexplore.ieee.org/document/8451714/ doi: 10.1109/ICIP.2018.8451714

Mirza, M., & Osindero, S. (2014). *Conditional Generative Adversarial Nets.* arXiv. Retrieved 2023-02-09, from http://arxiv.org/abs/1411.1784 (arXiv:1411.1784 [cs, stat])

Mu, B., Qin, B., Yuan, S., & Qin, X. (2020). A Climate Downscaling Deep Learning Model considering the Multiscale Spatial Correlations and Chaos of Meteorological Events. *Mathematical Problems in Engineering*, *2020*, e7897824. Retrieved 2021-11-20, from https://www.hindawi.com/journals/mpe/2020/7897824/ (Publisher: Hindawi) doi: 10.1155/2020/7897824

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., … Anandkumar, A. (2022). *FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators.* arXiv. Retrieved 2023-01-06, from http://arxiv.org/abs/2202.11214 (arXiv:2202.11214 [physics])

Price, I., & Rasp, S. (2022). Increasing the accuracy and resolution of precipitation forecasts using deep generative models. *arXiv:2203.12297 [cs, stat]*. Retrieved 2022-04-06, from http://arxiv.org/abs/2203.12297 (arXiv: 2203.12297)

Pulkkinen, S., Nerini, D., Pérez Hortal, A. A., Velasco-Forero, C., Seed, A., Germann, U., & Foresti, L. (2019). Pysteps: an open-source Python library for probabilistic precipitation nowcasting (v1.0). *Geoscientific Model Development*, *12*(10), 4185–4219. Retrieved 2023-01-18, from https://gmd.copernicus.org/articles/12/4185/2019/ doi: 10.5194/gmd-12-4185-2019

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., … Mohamed, S. (2021). Skilful precipitation nowcasting using deep generative models of radar. *Nature*, *597*(7878), 672–677. Retrieved 2022-04-20, from https://www.nature.com/articles/s41586-021-03854-z doi: 10.1038/s41586-021-03854-z

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204. Retrieved 2023-02-13, from http://www.nature.com/articles/s41586-019-0912-1 doi: 10.1038/s41586-019-0912-1

Roberts, N. (2008). Assessing the spatial and temporal variation in the skill of precipitation forecasts from an NWP model. *Meteorological Applications*, *15*(1), 163–169. Retrieved 2022-12-01, from https://onlinelibrary.wiley.com/doi/abs/10.1002/met.57 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/met.57) doi: 10.1002/met.57

Roberts, N. M., & Lean, H. W. (2008). Scale-Selective Verification of Rainfall Accumulations from High-Resolution Forecasts of Convective Events. *Monthly Weather Review*, *136*(1), 78–97. Retrieved 2022-12-01, from https://journals.ametsoc.org/view/journals/mwre/136/1/2007mwr2123.1.xml (Publisher: American Meteorological Society Section: Monthly Weather Review) doi: 10.1175/2007MWR2123.1

Saito, M., Matsumoto, E., & Saito, S. (2017). Temporal Generative Adversarial Nets with Singular Value Clipping. In *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 2849–2858). (ISSN: 2380-7504) doi: 10.1109/ICCV.2017.308

Seneviratne, S., Zhang, X., Adnan, M., Badi, W., Dereczynski, C., Di Luca, A., … Zhou, B. (2021). Weather and Climate Extreme Events in a Changing Climate. In V. Masson-Delmotte et al. (Eds.), *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 1513–1766). Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press. (Type: Book Section) doi: 10.1017/9781009157896.013

Serifi, A., Günther, T., & Ban, N. (2021). Spatio-Temporal Downscaling of Climate

Data Using Convolutional and Error-Predicting Neural Networks. *Frontiers in Climate*, *3*. Retrieved 2022-06-09, from https://www.frontiersin.org/article/10.3389/fclim.2021.656479

Sinclair, S., & Pegram, G. G. S. (2005). Empirical Mode Decomposition in 2-D space and time: a tool for space-time rainfall analysis and nowcasting. *Hydrology and Earth System Sciences*, 11.

Sun, A. Y., & Tang, G. (2020). Downscaling Satellite and Reanalysis Precipitation Products Using Attention-Based Deep Convolutional Neural Nets. *Frontiers in Water*, *2*. Retrieved 2022-03-09, from https://www.frontiersin.org/article/10.3389/frwa.2020.536743

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1–9). (ISSN: 1063-6919) doi: 10.1109/CVPR.2015.7298594

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning Spatiotemporal Features with 3D Convolutional Networks. In *2015 IEEE International Conference on Computer Vision (ICCV)* (pp. 4489–4497). (ISSN: 2380-7504) doi: 10.1109/ICCV.2015.510

Vandal, T., Kodra, E., Ganguly, S., Michaelis, A., Nemani, R., & Ganguly, A. R. (2017). DeepSD: Generating High Resolution Climate Change Projections through Single Image Super-Resolution. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1663–1672). Halifax NS Canada: ACM. Retrieved 2022-02-01, from https://dl.acm.org/doi/10.1145/3097983.3098004 doi: 10.1145/3097983.3098004

Vaughan, A., Tebbutt, W., Hosking, J. S., & Turner, R. E. (2022). Convolutional conditional neural processes for local climate downscaling. *Geoscientific Model Development*, *15*(1), 251–268. Retrieved 2022-02-15, from https://gmd.copernicus.org/articles/15/251/2022/ (Publisher: Copernicus GmbH) doi: 10.5194/gmd-15-251-2022

Vondrick, C., Pirsiavash, H., & Torralba, A. (2016). Generating Videos with Scene Dynamics..

Wang, F., Tian, D., Lowe, L., Kalin, L., & Lehrter, J. (2021). Deep Learning for Daily Precipitation and Temperature Downscaling. *Water Resources Research*, *57*(4), e2020WR029308. Retrieved 2021-11-20, from https://onlinelibrary.wiley.com/doi/abs/10.1029/2020WR029308 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2020WR029308) doi: 10.1029/2020WR029308

Wang, X., Lucas, A., Lopez-Tapia, S., Wu, X., Molina, R., & Katsaggelos, A. K. (2019). Spatially Adaptive Losses for Video Super-resolution with GANs. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1697–1701). (ISSN: 2379-190X) doi: 10.1109/ICASSP.2019.8682742

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., … Loy, C. C. (2019). ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In L. Leal-Taixé & S. Roth (Eds.), *Computer Vision – ECCV 2018 Workshops* (pp. 63–79). Cham: Springer International Publishing. doi: 10.1007/978-3-030-11021-5_5

Winterrath, T., Brendel, C., Hafer, M., Junghänel, T., Klameth, A., Lengfeld, K., … Becker, A. (2018). *Radar climatology (RADKLIM) version 2017.002; gridded precipitation data for Germany: Radar-based quasi gauge-adjusted five-minute precipitation rate (YW).* Deutscher Wetterdienst (DWD). Retrieved 2022-12-21, from https://opendata.dwd.de/climate_environment/CDC/help/landing_pages/doi_landingpage_RADKLIM_RW_V2017.002-en.html (Artwork Size: approx. 500 MByte per gzip compressed ascii or binary archive with monthly data Medium: gzip compressed and tar packed ascii and binary Pages: approx. 500 MByte per gzip compressed ascii or binary archive with monthly data Version Number: 1 Type: dataset) doi: 10.5676/DWD/RADKLIM_YW_V2017.002

Winterrath, T., Brendel, C., Hafer, M., Junghänel, T., Klameth, A., Walawender, E., … Becker, A. (2017). *Erstellung einer radargestützten Niederschlagsklimatologie* (Tech. Rep. No. 251). Deutscher Wetterdienst. Retrieved 2022-12-21, from https://www.dwd.de/DE/leistungen/pbfb_verlag_berichte/pdf_einzelbaende/251_pdf.pdf?__blob=publicationFile&v=2

spateGAN: Spatio-Temporal Downscaling of Rainfall Fields using a cGAN Approach

Luca Glawion[1], Julius Polz[1], Harald Kunstmann[1,2], Benjamin Fersch[1], Christian Chwala[1,2]

[1]Institute of Meteorology and Climate Research, Karlsruhe Institute of Technology, Campus Alpin, Garmisch-Partenkirchen, Germany
[2]Chair of Regional Climate and Hydrology, Institute of Geography, University of Augsburg, Augsburg, Germany

Key Points:

- High performance simultaneous spatial and temporal precipitation downscaling enabled by 3D convolution approach
- Generation of realistic high-resolution ensembles using probabilistic conditional generative adversarial networks
- Low computational effort compared to dynamical downscaling approaches

Corresponding author: Luca Glawion, `luca.glawion@kit.edu`

Abstract

Climate models face limitations in their ability to accurately represent highly variable atmospheric phenomena. To resolve fine-scale physical processes, allowing for local impact assessments, downscaling techniques are essential. We propose spateGAN, a novel approach for spatio-temporal downscaling of precipitation data using conditional generative adversarial networks. Our method is based on a video super-resolution approach and trained on ten years of country wide radar observations for Germany. It simultaneously increases the spatial and temporal resolution of coarsened precipitation observations from 32 km to 2 km and from 1 hour to 10 minutes. Our experiments indicate that the ensembles of generated temporally consistent rainfall fields are in high agreement with the observational data. Spatial structures with plausible advection were accurately generated. Compared to trilinear interpolation and a classical convolutional neural network, the generative model reconstructs the resolution-dependent extreme value distribution with high skill. It showed a high Fractions Skill Score of 0.73 for rainfall intensities over $15\,mm\,h^{-1}$ and a low BIAS of 3.55%. A power spectrum analysis confirmed that the probabilistic downscaling ability of our model further increased its skill. We observed that neural network predictions may be interspersed by recurrent structures not related to rainfall climatology, which should be a known issue for future studies. We were able to mitigate them by using an appropriate model architecture and model selection process. Our findings suggest that spateGAN offers the potential to complement and further advance the development of climate model downscaling techniques, due to its performance and computational efficiency.

Plain Language Summary

Natural disasters like floods, hail, or landslides originate from precipitation. Global climate models are an important tool to understand these hazards and derive expected changes in a future climate. However, they operate on spatial and temporal scales that limit the regional ability to reflect their small scale characteristics. This has led to the development of dynamical and statistical downscaling methods. Due to their computational efficiency, machine learning algorithms recently get increased attention as method for improving the spatial resolution of climate data. Here, we describe a new deep learning model that allows to simultaneously increase both the temporal and spatial resolution of precipitation data. Our presented approach enhances the spatial resolution by a factor of 16 and the temporal resolution by factor of 6. The generated rain fields are hardly identifiable as artificial generated and exhibit the typical structure, movement and distribution of observed rain fields.

1 Introduction

In the 2010s around 83 % of all natural disasters were caused by weather and climate extremes killing more than 410,000 people. Half of all disasters were a direct consequence of precipitation extremes like floods or landslides (IFRC, 2021). Rising average temperatures are expected to further increase both mean and extreme precipitation (Seneviratne et al., 2021), a development that may even be underestimated in climate projections (Allan & Soden, 2008). In order to adapt to a changing climate, accurate local and global information about the current and future hydrological cycle is indispensable. However, precipitation shows high spatial and temporal variability, exhibiting fluctuations on almost all spatial and temporal scales (Berg et al., 2013). Dynamical global climate models are restricted to larger scales by their high computational demand and for numerical stability criteria. With typical horizontal grid spacing of 30–80 km (Chen et al., 2021) and temporal resolutions of 1–24 hours, they are beyond of resolving fine-scale physical processes, extreme precipitation in particular. Due to subgrid-scale parameterizations, conclusions about the development of small-scale processes under a changing climate are not generally limited. However, for physically-based local climate impact studies, the characterization of high-resolution information about precipitation and its extremes is inevitable.

Consequently, downscaling methods have been developed and applied to increase the resolution of climate model outputs. These methods include statistical and dynamical downscaling using regional climate models, as well as AI-based downscaling that leverages artificial neural networks (ANNs), which have become increasingly popular in recent years. The AI-based downscaling methods are based on the image "super-resolution" approach which originates from computer science, precisely computer vision, where the resolution of optical images is increased (Dong et al., 2016; Kim et al., 2016; J. Johnson et al., 2016). The logical extension of this approach to the temporal domain is called "video-super-resolution" (Lucas et al., 2018; X. Wang, Lucas, et al., 2019). While the original application of super-resolution is based on a clear understanding of the data-generating process, the processes of generating climate observations are less well understood, presenting both a challenge and an opportunity for the application of ANNs (Reichstein et al., 2019). Following the super-resolution approach, high-resolution observational, climate model, or reanalysis data are first spatially coarsened to a lower resolution. The training objective of the ANN is to recover the original resolution. For example, in precipitation downscaling, high-resolution weather radar observations enable the modeling of complex precipitation patterns using ANNs. An additional benefit of ANNs is a considerable reduction in computation time and energy compared to traditional dynamical models (Pathak et al., 2022).

First approaches for spatial precipitation downscaling with ANNs used a deterministic convolutional neural network (CNN) which does not account for potential biases between observations and global climate model data or cover uncertainties related to the highly underdetermined problem (Vandal et al., 2017; F. Wang et al., 2021). Recent studies have extended the spatial super-resolution approach to the temporal domain and generated a single image with a fourfold higher spatio-temporal resolution applied to rainfall and temperature data (Serifi et al., 2021). CNNs have also shown their potential in downscaling low-resolution climate model outputs while outperforming other statistical approaches (Baño-Medina et al., 2020; Mu et al., 2020; Sun & Tang, 2020; Vaughan et al., 2022).

Recently, conditional generative adversarial networks (cGANs) (Mirza & Osindero, 2014) have been becoming increasingly popular for data generation problems. In comparison to classical CNN approaches, their advantages are that they do not rely on a predefined expert metric, but instead utilize an evolving metric in the form of an individual trained neural network. Furthermore, they have a stochastic design which enables them to generate an ensemble of solutions (Goodfellow et al., 2014). cGANs consist of two networks: a generator and a discriminator. The generator, typically a CNN, generates high-resolution images conditioned on low-resolution inputs, whereas the discriminator evaluates the quality of the generated images by distinguishing between real and artificial images. The generator´s task trying to trick the discriminator is defined by the model´s objective function (Ledig et al., 2017; X. Wang, Yu, et al., 2019). Both networks are simultaneously trained in an adversarial manner. This concept of a two-part architecture and model training has increased the generative performance of neural networks significantly, which is illustrated by the creation of realistic human faces (Karras et al., 2019). In climate science, cGANs can learn to reconstruct high-resolution solutions from climate model outputs and random components. Leinonen et al. (2021) demonstrated the performance and capability of cGANs within a spatial super-resolution approach by downscaling coarsened precipitation data from a resolution of 16 km to 1 km. The same idea has also been applied to downscaling global precipitation forecasts (Price & Rasp, 2022; L. Harris et al., 2022). Furthermore, cGANs outperformed traditional precipitation nowcasting algorithms (Ravuri et al., 2021).

Mapping low- to high-resolution precipitation data is an underdetermined problem due to fluctuations across scales. Resolving the temporal evolution of precipitation events in terms of intensity and advection, is necessary to obtain a complete picture of the high variability of precipitation and the expression of extreme events. Kashinath et al. (2021) refer to the generation of spatially and temporally coherent fields as the holy grail of downscaling. However, existing deep learning methods for spatio-temporal downscaling using CNN based downscaling methods can not sufficiently represent the high variability of precipitation due to their deterministic nature. Even though cGANs have proven to be suitable to present a probabilistic solution for the problem, the focus so far has been on increasing spatial resolutions without temporal downscaling. Often, the super-resolution approaches also address spatial or temporal scales not directly transferable to global climate model data. Furthermore, "recurrent structures" such as reappearing local biases in the generated fields can be an issue. This will also be addressed later in this manuscript.

In this study we propose spateGAN, a cGAN for spatio-temporal downscaling of precipitation based on the video super-resolution approach. We compare a deterministic version of the model to a probabilistic version. Precisely, the objective of this study is:

1. To evaluate the ability of a 3D fully-convolutional cGAN to simultaneously downscale rainfall fields in space and time, from a spatial resolution of 32 km to 2 km and temporally from 1 hr to 10 min.
2. To analyze the model results with respect to spatial structures, temporal consistency and extreme value statistics of the generated fields.

2 Methods

In the following we introduce a new spatio-temporal downscaling approach using a conditional generative adversarial network that learned to downscale spatially and temporally coarsened gridded precipitation observations from a weather radar network (Figure 1). As an evaluation case study we applied the final trained models to the domain of whole Germany and a time period consisting of 12 weeks of data distributed over all seasons. We compared a deterministic and a probabilistic cGAN (spateGAN$_{det}$ and spateGAN$_{prob}$) to a classical CNN approach and trilinear interpolation.

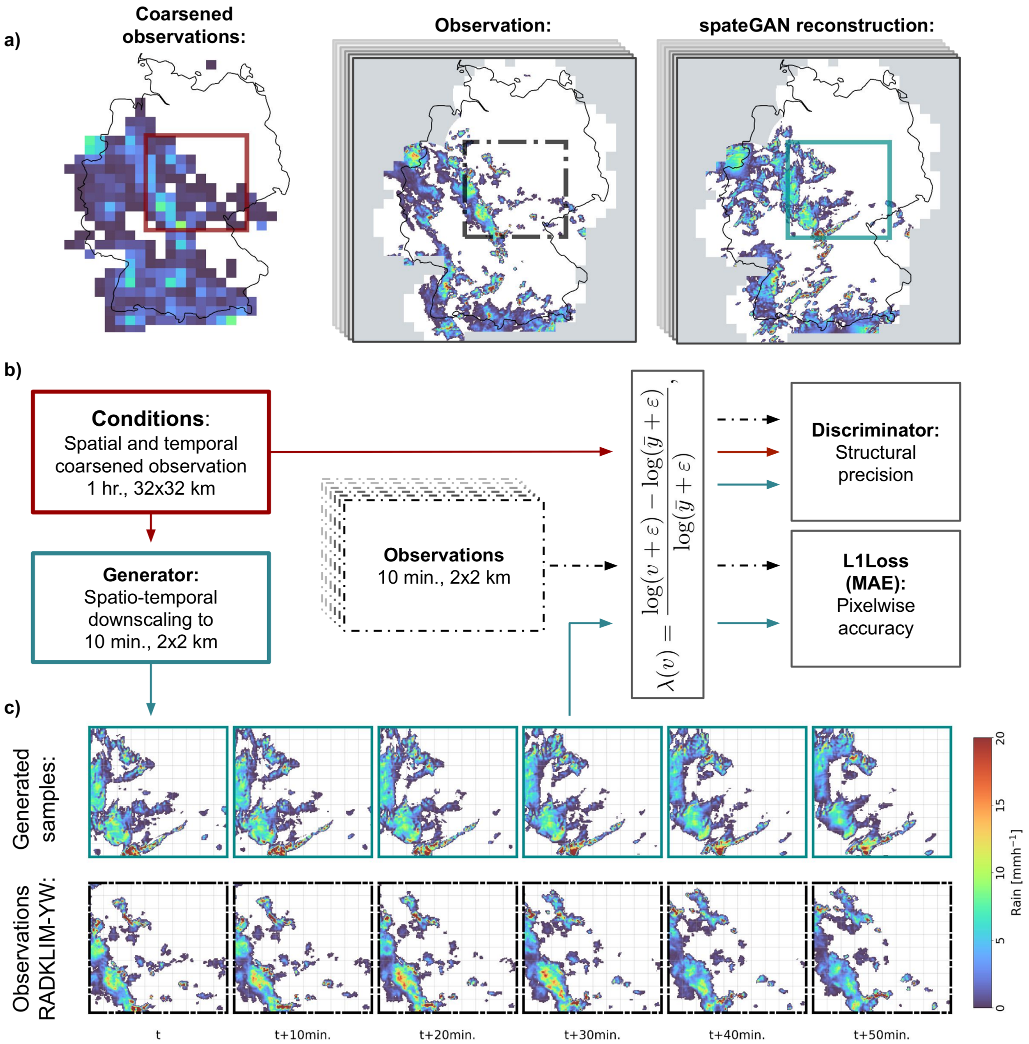

Figure 1. Overview of the proposed spateGAN model for spatio-temporal downscaling of precipitation data. The Figure illustrates the downscaling of a complex precipitation event in Germany, with both stratiform and convective elements. (a) spateGAN downscales coarsened data, derived from weather radar images, with arbitrary spatial and temporal dimensions from a resolution of 32x32 km and 1 hour to a higher resolution of 2x2 km and 10 minutes. The model is trained on smaller patches, represented by the colored boxes. (b) Schematic overview of the model components and training process. (c) Detailed downscaling results from a). spateGAN$_{det}$ is able to convert the hourly resolved coarsened data into a sequence of temporally consistent, finely structured precipitation fields, while also reconstructing the original distribution with higher precipitation intensities.

2.1 Conditional Generative Adversarial Networks for Downscaling

A conditional generative adversarial network comprises two neural networks, the generator G and the discriminator D, which are trained in an adversarial manner. G is a function

$$\begin{array}{rcl} G : \mathbb{R}^{t\times n\times m} & \rightarrow & \mathbb{R}^{d_t t\times d_s n\times d_s m} \\ x & \mapsto & G(x) \end{array} \tag{1}$$

that performs the actual spatio-temporal downscaling of the coarse input x by increasing the temporal resolution by a factor $d_t \in \mathbb{N}$ and the spatial resolution by a factor $d_s \in \mathbb{N}$. In this study $d_t = 6$ and $d_s = 16$. The number of time steps t and grid cells n, m were fixed during training, but can be larger during inference. The discriminator D is a classifier

$$\begin{array}{rcl} D : \mathbb{R}^{t\times n\times m} \times \mathbb{R}^{d_t t\times d_s n\times d_s m} & \rightarrow & \mathbb{R} \\ (x, y) & \mapsto & b \end{array} \tag{2}$$

that distinguishes whether the sequence of high-resolution rainfall maps y has been artificially generated from x (i.e. $y = G(x)$) or is the original high-resolution radar image corresponding to x (Figure 1, b). Both functions are defined as convolutional neural networks (see Section 2.2) trained in a so called adversarial training process. G and D improve their abilities, the generation and discrimination of realistic rainfall time sequences by alternatively minimizing and maximizing the objective function described in Section 2.3. The key point is the custom trainable objective function for G which does not require prior knowledge about the problem to be constructed, but is learned from the data itself via D. The data set and its preparation is explained in Section 2.5. The selection of an optimal model during training and its evaluation requires metrics that we introduce in Section 2.6.

Opposed to the downscaling task is the coarsening operator that was used to synthetically produce coarsened data from high-resolution images. We can define it by

$$\begin{array}{rcl} C : \mathbb{R}^{d_t t\times d_s n\times d_s m} & \rightarrow & \mathbb{R}^{t\times n\times m} \\ y & \mapsto & C(y), \end{array} \tag{3}$$

where $C(y)_{i,j,k} := \frac{1}{d_t d_s^2} \sum_{i'=i}^{i+d_t} \sum_{j'=j}^{j+d_s} \sum_{k'=k}^{k+d_s} y_{i',j',k'}$ is the average over d_t time steps and d_s by d_s grid cells. If not mentioned otherwise we will refer to y as the original high-resolution observation image that was used to produce x, i.e. $x = C(y)$.

2.2 Network Architecture

G and D are convolutional neural networks with a model architecture (Figure 2 a) built from three principal functional blocks (Figure 2 b). G is fully convolutional. The final architecture resulted from an iterative model optimization with special focus on spatio-temporal consistency and the absence of recurrent structures and artifacts. Due to the training time of several days, a full hyperparameter tuning routine and ablation study had to be omitted. For both networks we included 3D convolutional layers. For D these allow the extraction of spatio-temporal features of rain field structures for the decision making. For G they allow to account for spatial and temporal non-linear correlation embedded in the given conditions (Tran et al., 2015) and the reconstruction of temporally consistent high-resolution rainfall fields.

Convolutional-Block

The *Convolutional-Block* is intended to efficiently represent spatio-temporal structures within a feature map. The first part processes the input data through a 3D convolutional layer with kernel size $1 \times 1 \times 1$. Depending of the previous layer, the feature dimensionality is decreased to save computational costs and allow for a deeper model (Szegedy et al., 2015). This is followed by a ReLU activation function, another 3D-convolutional layer with kernel size $3 \times 3 \times 3$, a Batch Normalization layer and another ReLU activation (Ioffe & Szegedy, 2015).

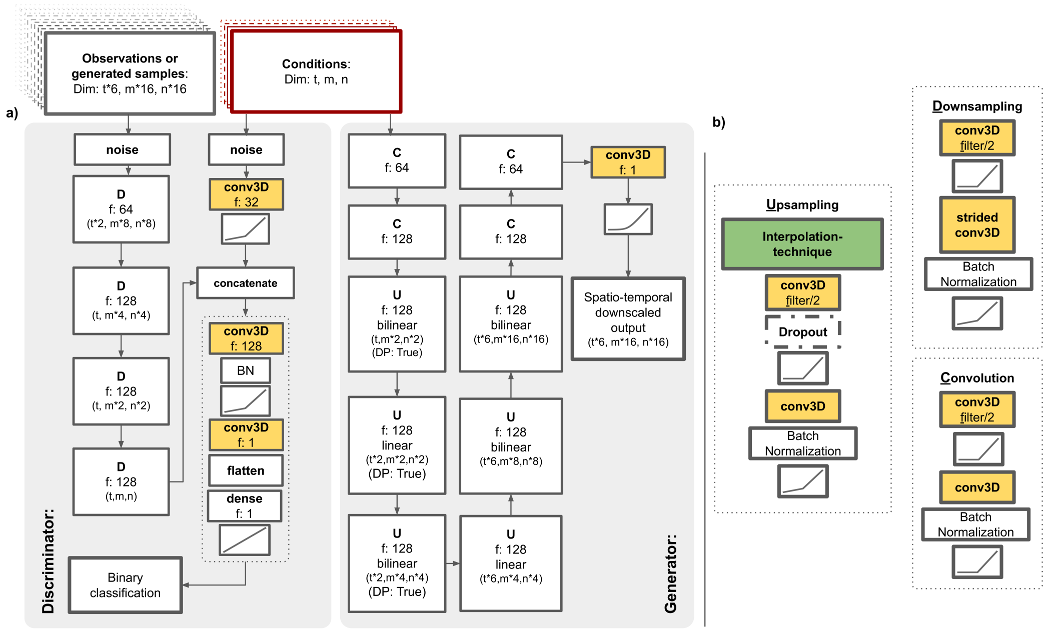

Figure 2. Detailed model architecture of spateGAN consisting of a generator and a discriminator. (a) The discriminator acts as a classification model, evaluating whether the high-resolution time sequences it receives are real or artificial, taking into account their possible affiliation with the coarsened input data provided as a condition. The generator spatially and temporally downscales the coarsened input data. For spateGAN$_{prob}$ dropout layer within the first three *Upsampling-Blocks* enable ensemble generation. (b) Architectures of *Upsampling, Downsampling and Convolutional Blocks*, the main components of both networks.

Upsampling-Block

The upsampling part of the network intends to increase the resolution of the input data by refining the grid size using bilinear interpolation in the spatial dimensions and linear interpolation for the time dimension. Each interpolation step is followed by a *Convolutional-Block* using a leaky ReLU activation to prevent the complete inactivity of these layers.

Downsampling-Block

The *Downsampling-Blocks* are only used within the discriminator. They are based on the presented *Convolutional-Blocks*, but with a kernel size of $4 \times 4 \times 4$ within the second 3D convolutional layer combined with strided convolution and leaky ReLU as second activation function. The approach is similar to Isola et al. (2017) and uses the spatial and temporal stride operation to reduce dimensionality of extracted features.

Generator

The generator initially consists of two *Convolutional-Blocks* without Batch Normalization. Subsequently, the spatial and temporal resolution of the hidden representation is increased using six *Upsampling-Blocks* to achieve the factors $d_t = 6$ and $d_s = 16$ to increase the temporal resolution of 1 hr to 10 min and the spatial resolution from 32 km to 2 km. Each interpolation step is followed by a *Convolutional-Block* to adjust spatio-temporal structures. There are two final *Convolutional-Blocks*, where the second block has no Batch Normalization. The model output is determined by a final convolutional layer to reduce the filter dimension. A softplus activation function limits the distribution of the output to positive values, which can be directly interpreted as rainfall intensity in mm/10 min. For each convolutional layer within G with a kernel size > 1 we applied a reflection padding strategy to reduce boundary errors.

Since downscaling is in general an underdetermined problem, the model uncertainty is closely related to the possible valid realizations of the high-resolution image. The capability of ensemble generation can provide additional valuable information. Leinonen et al. (2021) have shown that for pure spatial downscaling noise, passed as an additional generator feature, is suitable for ensemble generation. We compared a deterministic cGAN approach (spateGAN$_{det}$) to an alternative probabilistic approach (spateGAN$_{prob}$) for ensemble generation, exploiting dropout layers (Isola et al., 2017) within the first three generator *Upsampling-Blocks* during model training and inference. The dropout rate was set to 0.2 with temporal constant selected neurons for each individual ensemble member.

Discriminator

One challenge in training the discriminator is that the given data should be distinguished solely based on the temporal and spatial structures and the distribution. As a first model layer we add noise following a Gaussian distribution (mean=0, stddev=0.05) to the high- and coarse-resolution data to counteract a decision making based on a potential numerical inexactness of the generator while the real images are quantized and a perfect match for the coarse data.

There are two input branches to the network. The high-resolution data is processed by a series of four *Downsampling-Blocks*. The first one has no batch normalization layer. The extracted features are concatenated with the coarsened model input data, that passed through one 3D convolutional layer and a leaky ReLU activation function. After another 3D convolutional layer, Batch Normalization and a leaky ReLU activation function, the filter dimension is reduced using a last 3D convolutional layer. The resulting output is flattend and passed to a single dense layer using a linear activation function allowing for binary classification similar to Ravuri et al. (2021). We observed that Batch Normalization would not be required in all downsampling blocks to get to a similar model performance. However, they lead to a faster desirable model state during training (Ioffe & Szegedy, 2015).

2.3 Objective Function

We express the objective functions for spateGAN following Isola et al. (2017) combining Binary Cross Entropy with a L1 loss term. The L1 loss term or mean absolute error is a pixel-wise error that is only applied to the generator objective. It ensures that the generated rain fields remain close to the ground truth. However, the distribution of rainfall deviates strongly from prominent ANN image data sets. Common methods to achieve a well-performing model and a stable training in spite of this, are data logarithmization and normalization routines (L. Harris et al., 2022; Leinonen et al., 2021; Price & Rasp, 2022).

This, however, can amplify the generation of unrealistically high rainfall intensities in case of a model overestimation during inference or training and a potential necessity of a limitation of the value range in form of an activation function like *sigmoid* or *tanh*, or by a fixed allowed maximum value. In our opinion such a constraint would limit the model to perform well in a non-stationary system. Therefore, we present a new alternative approach using an updated objective function. We logarithmized and normalized data that enter the discriminator or were considered for the calculation of the L1 loss according to

$$\lambda(v) = \frac{\log(v + \varepsilon) - \log(\bar{y} + \varepsilon)}{\log(\bar{y} + \varepsilon)}, \tag{4}$$

where $\bar{y}$ is the maximum of the high-resolution pixel values of the training data set (see Section 2.5.2) and $\varepsilon = 10^{-3}$.

The generator, on the other hand, as visualized in Figure 1 b), was provided unmodified input data and also produced output values that follow the original distribution of the radar data set. The final objective function is

$$\begin{aligned}\mathcal{L}_{cGAN}(G, D) \quad &= \quad \mathbb{E}_{x,y}[logD(\lambda(x), \lambda(y))] + \\ &\quad\quad \mathbb{E}_{x}[log(1 - D(\lambda(x), \lambda(G(x))))] + \\ &\quad\quad \alpha\mathbb{E}_{x,y}[||\lambda(y) - \lambda(G(x))||_1]\end{aligned} \tag{5}$$

where G tries to minimize this objective and the adversarial D tries to maximize it. We set α to 20, to align the loss terms to a comparable range. For spateGAN$_{prob}$ we consulted one random ensemble member per training step during model training for loss calculation to save computational resources.

2.4 Comparison Models: Trilinear Interpolation and Convolutional Neural Network

As a baseline model we refined the grid size of the coarsened validation data correspondingly by a spatial factor of $d_s = 16$ and temporal $d_t = 6$ using trilinear interpolation. In addition, we compared the performance of the spateGANs with a classical neural network approach. For this purpose, we trained a CNN with the exact same architecture as the generator of spateGAN$_{det}$ (see Section 2.2) only applying L1 loss from [5] without D. The remaining training routine was unchanged.

2.5 Radar Data

For model training, testing and validation we used RADKLIM-YW, a publicly available gauge-adjusted and climatologically-corrected weather radar product provided by the German Meteorologic Service (DWD) that can be retrieved from Winterrath et al. (2018) . The radar composite contains information of 16 weather radars adjusted by approx. 1000 rain gauges homogeneously distributed throughout Germany. A detailed description of the radar data processing and correction can be found in Winterrath et al. (2017).

The grid extent is $900\,\text{km} \times 1100\,\text{km}$ with a resolution of $1\,\text{km} \times 1\,\text{km}$. The temporal resolution is 5 minutes, where each grid cell represents a 5 minute rainfall sum. Regions not covered by the 150 km measurement radii of the radars or missing measured values are marked with "NaNs". For our investigation we used data from 1 January 2010 until 31 December 2021. After downloading we transformed the binary data to a NetCDF format following Chwala and Polz (2021) to be able to easily handle the large amounts of data (1Tb/year).

To prevent information leakage and to validate the model's ability to generalize outside the training distribution, the data was split into three sets: 2010–2019 for training, 2020 for testing, and 2021 for validation. All presented results stem from the validation data set.

2.5.1 Data Preprocessing

Before network training, testing and validation, suitable data was selected, the downscaling factor was defined and the high-resolution samples were coarsened. The spatial resolution should increase 16-fold from $32 \times 32\,\text{km}$ to $2 \times 2\,\text{km}$ and the temporal resolution 6-folded from 1 hour to 10 minutes. The chosen scales are sufficient to simulate the downscaling of global climate model data, which can be provided with similar resolution and to be fine enough to reveal the high temporal and spatial variability of precipitation. A further increase of the resolution towards the original RADKLIM-YW data ($1 \times 1\,\text{km}$ and $5\,\text{min}$) would have exceeded our currently available computational resources in terms of GPU memory. Consequently, as a first preprocessing step, the data was spatially averaged and temporal aggregated to a 2 km and 10 minute resolution.

2.5.2 Training and Testing Sample Preparation

GPU memory limitation did not allow the usage of longer time series of whole maps of Germany for model training and testing. Therefore, we randomly selected samples with a spatio-temporal extent of 160×160 pixels and 36 time steps, i.e. $320\,\text{km} \times 320\,\text{km} \times 6\,\text{hr}$. This approach also reduces the risk of the model memorizing spatial dependencies and patterns in the data.

The rain intensity in the data follows a near-lognormal distribution and only about 5 % of the pixels of the radar composite contain precipitation, leading to a high imbalanced and skewed distribution which is difficult for training neural networks. The main issue is learning reasonable predictions for the minority class (J. M. Johnson & Khoshgoftaar, 2019). For rainfall this refers to rarely occurring events and high precipitation intensities. To overcome this problem a simple data augmentation routine was applied. This routine balances the distribution of the train and test samples, increasing the number of wet pixels and total amount of precipitation, and allowing the model to focus on relevant rain events. The data augmentation process selected only samples free of missing values, total precipitation (of all time steps and pixels) exceeding 1000 kg and with at least 100 kg/ 10min per time step for 2/3 of all time steps. To avoid a systematic bias due to the prevailing westerly wind flow influence in Germany, half of the chosen samples were rotated (90° or 270°) or mirrored (vertically or horizontally).

In total, 112,500 samples were randomly drawn for model training (y_{train}) and 1000 samples y_{test} for model testing during training. The test data was also used for model selection (see Section 2.8). As a final preprocessing step, coarsened versions $C(y_{train})$ and $C(y_{test})$ were calculated, resulting in a final model input shape during training ($t \times n \times m$) of 6 time steps and 10×10 pixels.

2.5.3 Validation Data

To validate the model performance, we utilized the fully convolutional architecture of G to downscale entire maps of Germany. This entails a future possible application of downscaling global climate model outputs over a larger domain than the training samples dimension, and the model's ability to generalize for this. To include all seasons and connected temporal sequences, while reducing data volume, we selected the first week of each month of 2021 for validation, resulting in 12,096 validation time steps.

We applied $C(y_{val})$ to derive the coarse validation data, ignoring missing values and setting completely empty coarsened pixels to zero. After model prediction, we masked the downscaled data to exclude pixels with NaN values in y_{val} and areas of coarsened pixels that were not entirely within the radar network coverage, but intersect with it. Additionally we excluded the first and last hour of individually predicted time steps to avoid temporal boundary errors. We applied this procedure to contain all available information in the coarsened data, but derive valid predictions only for those areas where no data is missing. Evaluation metrics were calculated for a cropped area of $370 \times 560\,\text{km}$ (highlighted in Figure 6) to further mitigate boundary effects.

The length of time sequences downscaled by G is mutable and only limited by GPU memory. Using a NVIDIA Tesla V100, G is able to predict 66 time steps of high-resolution maps ($66 \times 480 \times 480$) from 11 coarse precipitation maps ($11 \times 30 \times 30$) in one single processing step, taking 0.1 seconds. Successive predictions were made for contiguous time sequences of this size, resulting in 11,652 images. For spateGAN$_{prob}$ we calculated, according to Section 2.2, 5 ensemble members (spateGAN$_{prob01,02etc.}$) using fixed drop-out neurons for each member and a sixth member, spateGAN$_{prob06}$, in which the selected neurons were randomly changed for every prediction step, i.e. 6 hours. The aggregation of this mixed ensemble member represents the accumulated ensemble mean in this study.

2.6 Metrics

The high temporal and spatial complexity of precipitation makes it difficult to validate the results using a single metric. In addition, different users and decision makers have different requirements on the capabilities of a downscaling model. Thus, the evaluation of the results was carried out with a set of metrics considering different spatial scales and temporal aggregations. Additionally, a qualitative analysis was performed. For calculating the following metrics and for all shown results, we set observed (R_{ref}) and generated (R_{gen}) rain rates below $0.01\,mm\,h^{-1}$ to zero.

2.6.1 Fractions Skill Score

The Fractions Skill Score (FSS) is a spatial verification method to evaluate the performance of precipitation forecasts. It is a measure of the rainfall misplacement error with respect to a given spatial and temporal scale (N. Roberts, 2008; N. M. Roberts & Lean, 2008). A neighborhood of a pixel P contains all grid cells in a r by r square centered at P and T previous and following time steps. Let f_{ref} be the fraction of grid values larger than δ contained in a neighborhood averaged over all possible neighborhoods in an observed image. We define f_{gen} in the same way using the generated image. Then the FSS for δ, r and T is defined by

$$FSS = \frac{\overline{(f_{gen} - f_{ref})^2}}{\overline{f_{gen}^2} + \overline{f_{ref}^2}}, \tag{6}$$

where $\overline{f}$ denotes the average over all images in the data set. For ensemble predictions the fraction is given by the average fraction over all ensemble members. We computed the FSS for various combinations of thresholds δ and scales, r and T.

2.6.2 Radially Averaged Logarithmic Power Spectrum Density

We computed the radially averaged power spectral density ($RAPSD$) and temporal power spectrum density PSD_t to analyze spatial and temporal patterns independent of their location (D. Harris et al., 2001; Sinclair & Pegram, 2005). The $RAPSD$ of a single image was obtained through transforming its 2D power spectrum into a 1D power spectrum by radial averaging, as implemented in PYSTEPS (Pulkkinen et al., 2019). The pixel wise power spectrum along the time dimension is referred to as PSD_t. We calculated the RAPSD for single images ($RAPSD_{10}$), hourly aggregated images ($RAPSD_{60}$) and the accumulation of the entire evaluation data set $RAPSD_{aggr}$.

We compared the power spectrum density of the artificially generated rain fields with the analog measure derived from the observation data. First, we used $RAPSD_{10}$ to evaluate spatial patterns in terms of their frequency and amplitude. Second, we used PSD_t and $RAPSD_{60}$ to quantify the ability to generate temporally consistent fields. And third, we used $RAPSD_{aggr}$ to reveal if models produce recurrent structures (local biases) that sum up over time and are distinct from recurrent local structures in the reference data. An example of such structures is given in Figure 6.

2.6.3 Point Wise and Distribution Error

As a point wise error we computed the mean absolute error (MAE) given by

$$MAE = \overline{|R_{ref} - R_{gen}|}. \tag{7}$$

The continuous ranked probability score ($CRPS$) is a generalization of the mean absolute error and evaluates a probabilistic models predictive distribution against observed values (Gneiting & Raftery, 2007).

The relative $BIAS$ measures the average model error as a percentage of the mean observed rainfall and is given by

$$BIAS = \frac{\overline{R_{gen} - R_{ref}}}{\overline{R_{ref}}} * 100 \tag{8}$$

The Kolmogorov-Smirnov (KS) test measures the maximal distance between the cumulative distribution of observed and generated rainfall. It evaluates the modelled distribution independent of the spatial distribution of values. Because of the skewed distribution of rainfall this maximal distance is most often located at low rainfall intensities which limits conclusions about extreme values.

2.7 Model Training

Each model was trained for three days resulting in about 3×10^5 training steps using mixed precision. The optimization of the spateGANs followed a standard approach by alternating between one gradient descent step for D, followed by one step for G (Goodfellow et al., 2014) and counted as one training step of the spateGAN. We trained on randomly selected samples from the training data set on one Nvidia Tesla V100 GPU limiting batch size to 7. For gradient descent, Adam optimizer was chosen with a learning rate of 1×10^{-4} for G (momentum parameters: $\beta_1 = 0.0, \beta_2 = 0.999$) and 2×10^{-4} for D ($\beta_1 = 0.5, \beta_2 = 0.999$). Models were saved after every 500th training step to later select the best performing state. We implemented the ANNs and model optimization in a Python framework using TENSORFLOW (version: 2.6) (Developers, 2022).

2.8 Model Selection

We selected the best performing models (i.e. the optimal state of either CNN, spateGAN$_{det}$ and spateGAN$_{prob}$ during training) by downscaling the test data. We took the structural error of all generated images into account using both $RAPSD_{aggr}$ and the average $RAPSD_{10}$. We represent the RAPSD deviation by a single value by calculating the mean absolute error of the logarithmized RAPSDs of predicted and real images:

$$\sigma = \frac{1}{n} \sum_{i=1}^{n} |10 * log_{10}(RAPSD_{real}) - 10 * log_{10}(RAPSD_{predicted})| \tag{9}$$

Based on $RAPSD_{aggr}$, σ_{aggr} considers potential model artefacts in the form of recurrent structures and the model ability to reconstruct adequate rain sums for a longer time period. Based on $RAPSD_{10}$, σ_{10min} takes the models ability to generate rain fields with spatial structures of the right amplitudes and frequencies into account. To avoid too strong influence of boundary errors in this selection we excluded the outermost edge, corresponding to one coarse resolution pixel, for this calculation. Finally, the model minimizing $\sigma_{aggr} + \sigma_{10min}$ was selected.

3 Results

To evaluate the spatio-temporal downscaling performance we considered the models capability to reconstruct the target distribution from spatially and temporally coarsened input data and to generate rain fields that closely resemble the observations regarding spatial structure and temporal consistency.

3.1 Qualitative Analysis

We start with a qualitative analysis examining a detailed visualization of the sequences generated for three rain events. One is a convective case study scenario and the other two show a stratiform and a mixed type rain event. The observation data, their

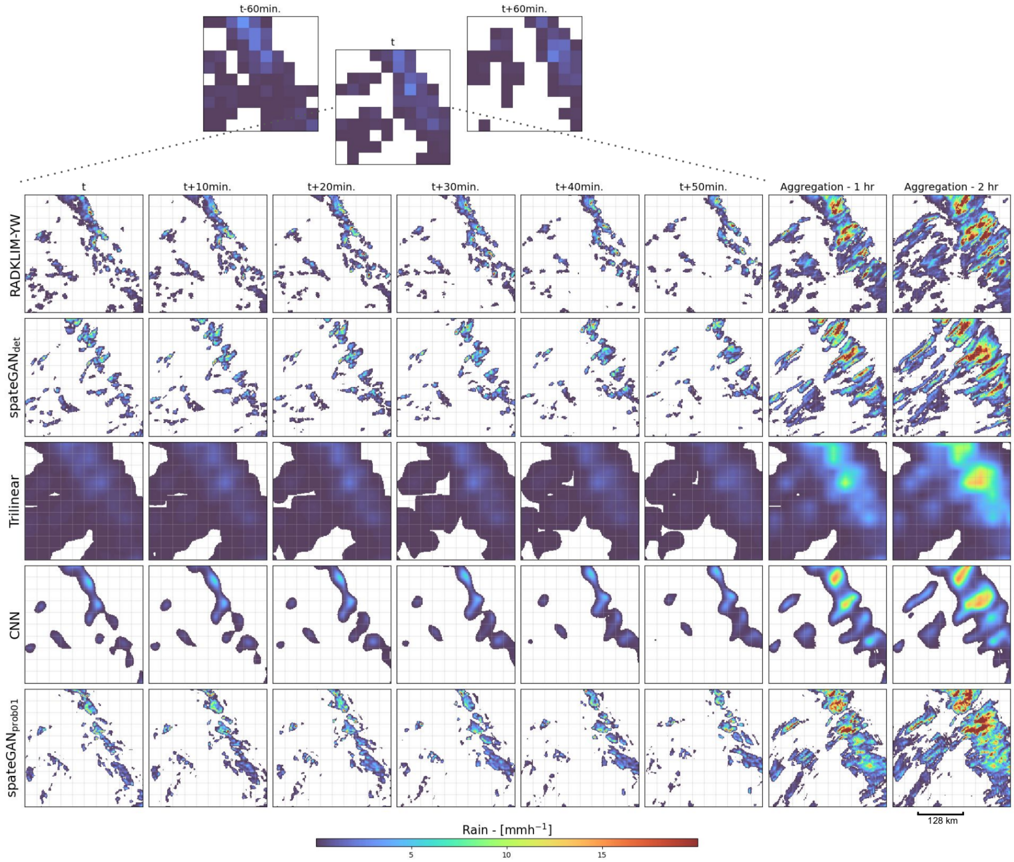

Figure 3. Detailed case study of the spatio-temporal downscaling performance for a convective precipitation event for central Germany. Shown are a temporal sequence of coarsened model input data, associated RADKLIM-YW observations, and model predictions. Hourly and two-hourly aggregated images highlight specific advection structures.

associated coarsened representation and the respective models are shown in Figures 3, 4 and A1. The predictions from the probabilistic generative approach stem from a single ensemble member (spateGAN$_{prob01}$). Additionally, the preceding and subsequent time steps of the coarsened images are presented to provide a better understanding of what information is available to the model to generate the high-resolution images. A more complete picture is given by the attached animations visualizing the full time sequences of different events (`https://doi.org/10.5281/zenodo.7636929`).

Case Study: Convective Rain Events

Figure 3 shows the temporal evolution of a convective rainfall event. The challenge for the downscaling models was to determine that the connected rainfall field in the coarsened input data represents disconnected convective cells and to localize them correctly with plausible advection.

Both spateGAN approaches effectively generated small convective rain cells from the low-resolution data which cannot be easily identified as artificially generated. The spatial structures, localization and advection were in good agreement with the observation data. However, there are differences in certain regions. For example, a more connected rain field in the north was represented as smaller separated cells. The observed small rain event in the southeast at t+20 min with a rain rate $> 20\, mm\, h^{-1}$ was generated as a larger event with lower rain rates. Despite these small scale dissimilarities, spateGAN was able to construct plausible local extremes like in the northern part of the images. In addition to the individual time steps, the 1-hour aggregations revealed advection structures that are very similar to the observation data in large parts of the images. This supports the hypothesis that the model is able to reproduce spatio-temporally consistent small-scale rainfall structures with plausible advection.

The CNN could generate rain fields with reasonable position and timing, but the cells lacked fine-scaled spatial structure and local extremes. Especially the gradients were very smooth. The model was not able to separate individual convective cells, however by comparing the presented time steps in chronological order, a plausible movement and temporal consistency became apparent.

The trilinear interpolation created a blurry version of the low-resolution data lacking local gradients, extreme values or advection.

Case Study: Stratiform Rain Events and Embedded Convection

Figure 4 presents the one hour time sequence of a stratiform rain event. The challenge for the models was to reconstruct the evolution of this larger rain field including areas with no precipitation and a smaller separated cell in the north, from contiguous pixels in the coarsened input data. The results from the spateGANs appear very similar to the observational data, including the size and positioning of the generated rain fields. The artificially generated events show plausible structures with a slight underestimation of the maximum rainfall intensity in, e.g., image t+20min. Higher rainfall intensities in the southeast corner and correctly positioned holes were created. The small detached rain events in the north are also depicted and are hardly distinguishable from the observation data. The generated structures exhibit a plausible temporal and spatial development, even though the rain field is moving slowly. spateGANs ability to generate both small and large rain events in a single image is further demonstrated for a complex precipitation event in Figure A1.

As within Figure 3, the trilinear interpolation and CNN results were blurry and lacked spatial structure. The CNN was more accurate in terms of the spatial extent of the rain field, while the trilinear interpolation produced fields that exceeded the spatial extent of the reference.

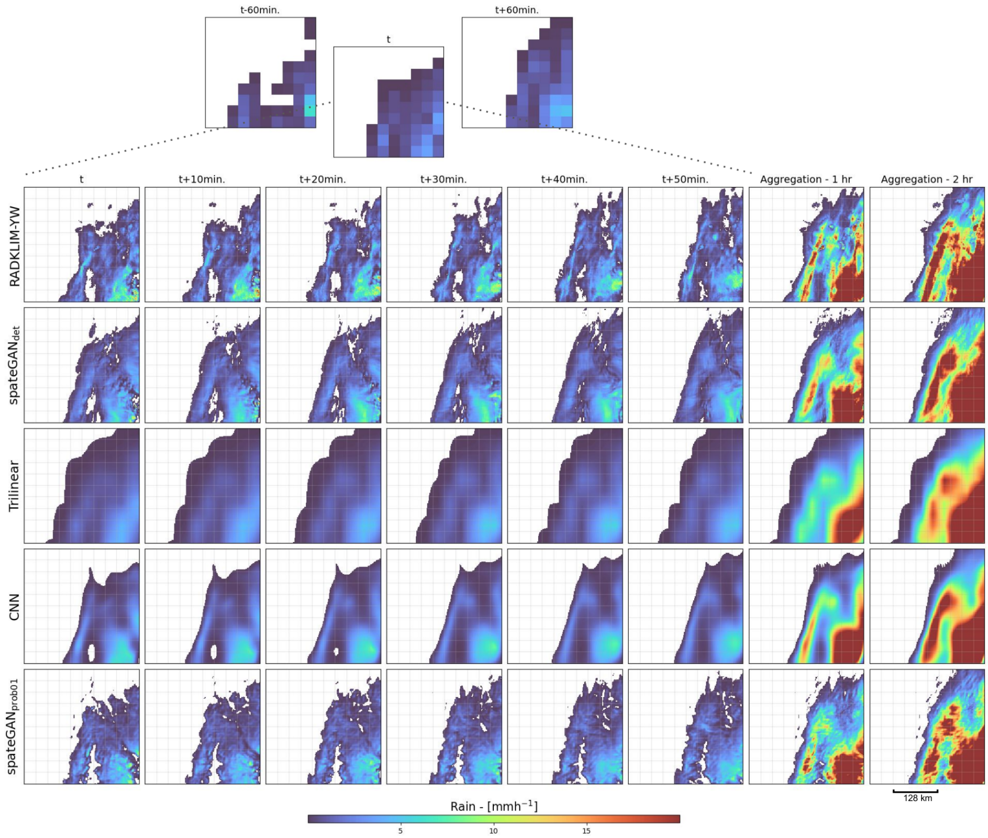

Figure 4. As Figure 3 for a stratiform event.

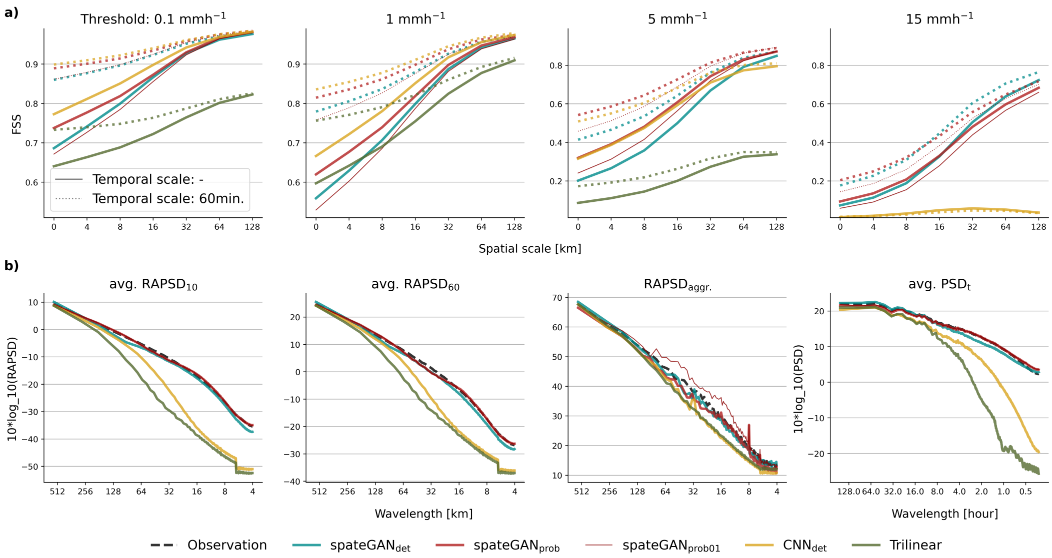

Figure 5. Evaluation of the downscaling methods (spateGANs, CNN, and trilinear interpolation) for a cropped area of the 2021 validation data set for Germany. (a) presents the Fractions Skill Score (FSS) for different thresholds and spatial and temporal scales, with the ensemble FSS of multiple members for spateGAN$_{prob}$. Part (b) evaluates the generated spatial and temporal structures using power spectra analysis. spateGAN$_{prob}$ refers not to multiple ensemble members, but to the mixed ensemble member as described in Section 2.5.3. The temporal consistency of the generated fields is evaluated using $RAPSD_{60}$ and the average PSD_t. All ANN models show peaks in $RAPSD_{aggr.}$ at different wavelengths and intensities, indicating the presence of recurrent patterns in the predictions.

3.2 Quantitative Investigation

The quantitative analysis is divided into two parts. First we investigated the models regarding their capability to generate detailed spatio-temporal rain field structures by analyzing the power spectrum. Then, we examined the pixel accuracy and the ability to reconstruct a skillful distribution in time and space by calculating the FSS, CRPS, MAE, KS statistics and BIAS.

3.2.1 Structural Analysis

We calculated the average $RAPSD_{10}$ and $RAPSD_{60}$ of the high-resolution observation images and the associated model predictions to investigate whether the models are able to represent the structural variability and advection of precipitation across spatial and temporal scales. The same analysis was performed for the accumulated precipitation of all 11652 validation images ($RAPSD_{aggr.}$) to visualize potential undesirable model characteristics such as the generation of recurrent structures that would manifest as peaks at certain wavelengths.

Figure 5 b) shows that the generated images from spateGAN$_{det}$ and spateGAN$_{prob}$ have a high structural similarity to the observations for both, single images and hourly aggregations on all considered scales. A small underestimation occurred between wavelengths of 128 to 64 and < 6 km for spateGAN$_{det}$. Respectively a slight overestimation occurred for spateGAN$_{prob}$. The same was observable in the temporal power spectrum PSD_t for wavelengths between 30 min. and 4 hours. For higher frequencies spateGAN$_{prob}$ showed a slight overestimation. The $RAPSD_{aggr}$ was close to the observation data. How-

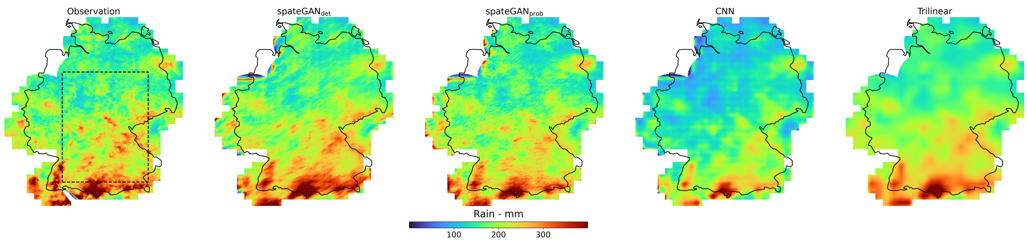

Figure 6. Aggregated observed and predicted rainfall of the validation data set for Germany for the year 2021. The accumulation shows the models ability to maintain the total rainfall amount and reveals recurrent structures within the predictions that contradict the physical principle of developing rain fields. spateGAN$_{prob}$ represents an ensemble mean as described in Section 2.5.3 and the rectangle defines the area considered for the quantitative analysis.

ever, peaks mainly prominent at a wavelength of 8 and 6 km could be observed. Recurrent structures with this frequency were also visible in the accumulated rainfall maps from Germany in Figure 6. Predictions of spateGAN$_{det}$ also exhibited this conspicuity at a wavelength of 32 km. At shorter aggregations (e.g. individual predictions, $RAPSD_{10}$ or $RAPSD_{60}$) these structures were not detectable.

For the CNN, $RAPSD_{10}$, $_{60}$ and $_{aggr.}$ showed an underestimation, especially for higher frequencies. This results from the missing model ability to generate small scale structures and to reconstruct the original high-resolution distribution. Recurrent structures could be also observed at wavelength of 32 km.

Trilinear interpolation was in general not capable to generate small scale spatio-temporal structures that were similar to the observation data. A high RAPSD and PSD_t underestimation could be shown for wavelength smaller 128 km or 8 hours. Within the whole accumulated validation data set no recurrent structures could be observed considering $RAPSD_{aggr}$ or Figure 6.

3.2.2 Distribution Reconstruction Skill

The coarse resolution provided as model input compresses the distribution of rainfall intensities towards lower values. The decisive factor of a skilful downscaling model is therefore not only the generation of realistic spatial structures, but rather the ability to reconstruct the correct distribution of rainfall intensities with accurate spatial and temporal placement of the rain events. We measured this downscaling skill by considering the FFS for the spatial and temporal precision of reconstructing high intensities using thresholds δ of 0.1 $mm\,h^{-1}$, 1 $mm\,h^{-1}$, 5 $mm\,h^{-1}$ and 15 $mm\,h^{-1}$. These thresholds represent the 0.9, 0.97, 0.997 and 0.9998 quantiles of the validation data set. The spatial scales r were between 0 and 128 km and the temporal scales T were 0 and 60 minutes. The results are shown in Figure 5 a). The generative models demonstrated a high skill for small to moderate rainfall (0.1 and 1 $mm\,h^{-1}$) with FSS exceeding 0.9 at a spatial scale of 32 km. They also performed well for high and strong rainfall intensities, with FSS values over 0.8 and 0.7 for a threshold of 5 and 15 $mm\,h^{-1}$. The score of spateGAN$_{prob}$ increased further, especially for small rain rates and scales, when multiple ensemble members were considered and the ensemble FSS was calculated. The CNN showed the best performance for small and moderate rainfall rates, but the accuracy decreased for strong rainfall intensities with a maximum FSS of 0.06 for 15 $mm\,h^{-1}$. Trilinear interpolation performed well for moderate precipitation (1 $mm\,h^{-1}$) but had the lowest overall skill.

Additionally, we calculated pixel accuracy metrics CRPS, or MAE for deterministic models, and the BIAS, as well as the distribution error as the KS statistics shown in Table 1. In terms of MAE, KS statistics, and BIAS the spateGAN models achieved

Table 1. Set of downscaling skill metrics computed for the validation data set. The FSS refers to the maximum score of Figure 5 a) each model achieved for different thresholds. For spateGAN$_{prob}$ multiple ensembles were considered for CRPS and FSS, a single member for MAE, KS statistic, power spectra deviation σ_{10min} [9] and BIAS.

	CRPS/MAE	KS	FSS$_{0.1}$	FSS$_1$	FSS$_5$	FSS$_{15}$	σ_{10min}	BIAS
spateGAN$_{det}$:	-/0.018	0.010	**0.98**	0.97	0.87	**0.73**	1.36	3.35
spateGAN$_{prob}$:	**0.012**/0.018	0.014	**0.98**	0.97	**0.89**	0.71	**0.31**	-3.55
CNN:	-/**0.012**	**0.008**	**0.98**	**0.98**	0.81	0.06	16.1	-22.22
Trilinear:	-/0.016	0.20	0.81	0.91	0.23	0	18.6	**-0.25**

overall good scores, compared to CNN and trilinear interpolation. The BIAS of spateGAN$_{det}$ showed a slight overestimation and an underestimation for spateGAN$_{prob}$. The CNN had the best KS score and MAE, but a negative BIAS of -22.28 % indicated a strong underestimation (see Figure 6). Trilinear interpolation showed the best BIAS with -0.28 % .

3.3 Ensemble Downscaling

The generation of multiple ensemble members is crucial to quantify uncertainties in the downscaling process like the likelihood of extreme events (Pathak et al., 2022).

By comparing the probabilistic generative approach to the deterministic, it could be shown that the predictions of an individual ensemble member, like spateGAN$_{prob01}$, looked similarly realistic as the predictions of spateGAN$_{det}$ (see Figure 3, 4 and A1). Regarding the RAPSD$_{10}$, RAPSD$_{60}$ and PSD$_t$ the predictions where even closer to the observation data as can be seen in Figure 5. The downscaling skill of spateGAN$_{prob01}$ was only minimally reduced with lower FSS for the thresholds 0.1, 1 and 15 $mm\,h^{-1}$, but higher scores for 5 $mm\,h^{-1}$. The potential of a probabilistic approach which considers multiple spateGAN$_{prob}$ ensemble members was investigated by calculating the CRPS and ensemble FSS (see Table 1). The CRPS showed an improvement with a value of 0.012 compared to the MAE of SpateGANdet and SpateGAN$prob$01. Furthermore, the FSS indicated a better downscaling performance compared to SpateGANdet and SpateGAN$prob$01, particularly for small scales and low rainfall amounts. The probabilistic model was also able to well represent the precipitation sum of the validation reference considering the aggregated ensemble mean, as can be seen in Figure 6.

However, Figure 5 shows that the aggregation of a single ensemble member (RAPSD$_{aggr}$ for spateGAN$_{prob01}$) showed an overestimation from scales between 8 and 128 km. We assume that this model characteristic was due to the chosen dropout routine. For one ensemble member selected drop out neurons were fixed for all time steps. The behaviour was not visible in single predictions and could only be revealed via the aggregation and analysis of multiple thousand images. To address this constraint, we emphasize to always consider multiple ensemble members, when applying this approach for longer time series.

Furthermore, we experimented to change the drop out rate after model training, which lead to an increased variance of the ensemble members. However, the downscaling skill was not further improved. Additionally, we trained a model applying random drop out neurons for each time step, which could generate temporal consistent rain fields without issues when aggregating single ensemble members. However, it frequently produced low rain rates during dry time steps and regions. Overall this exemplifies that various approaches for ensemble generation are feasible, but the creation of ensembles that reflect the physical plausible solutions and the stochasticity of the target data set is challenging and therefore subject to further research.

4 Discussion

In this study we proposed spateGAN, a novel approach for spatio-temporal downscaling of precipitation data combining cGANs, 3D convolution and interpolation techniques. It effectively increases the spatial resolution of coarsened weather radar data from 32 km x 32 km and 1 hour to 2 km x 2 km and 10 minutes. In the following we will discuss the models ability to accurately reconstruct spatial structures with temporal consistency and correct extreme value statistics. Additionally, we present the models limitations and additional unexpected findings.

Spatial Structures

The qualitative investigation (see Section 3.1) and the presented animation prove the ability of spateGAN to generate plausible precipitation fields from coarsened input data that are hardly classifiable as artificially generated. This is supported by the power spectrum analysis using $RAPSD$ and PSD, which are in highest agreement with the observation data for all scales when compared to CNN and interpolation. The FSS confirms that unlike trilinear interpolation and a classical CNN approach, the cGAN approach accurately produces structures with higher rainfall intensities. spateGAN is the only model that is able to generate rain cells of small spatial extent (see Figure 3). Besides the spatial extent and the rainfall intensity, the number of generated cells has a similar order of magnitude compared to the observations. Only the precise location of these cells deviates due to the stochastic nature of the model. spateGAN also tends to produce slightly smoother structures than the observed ones for large scale rain events like shown in Figure 4. We assume that an increase of the training sample dimensions could improve the structural quality of such large rain events. Overall, the results emphasize the necessity of a generative network downscaling approach for modeling realistic rain fields, since trilinear interpolation and CNN lack higher frequencies in the power spectrum. Trilinear interpolation approximates the low-resolution data providing limited additional information, while the CNN generates more detailed, but still too blurry events (Larsen et al., 2016).

Temporal Consistency

The animations of downscaled rain fields illustrate temporal consistency as a key property of spateGAN. The generated fields exhibit plausible advection, showing that rain cells are not randomly appearing and disappearing between time steps. This is supported by the 1 hour and 2 hour aggregations (see case study Figures 3, 4, A1), where the sum of individual time steps leads to smooth, connected cells elongated in the direction of advection. Furthermore, $RAPSD_{60}$ and PSD_t are in high agreement with the observation data. The visual evaluation of the CNN predictions and its improved PSD_t compared to trilinear interpolation also indicate the CNN's ability to generate temporally consistent events. This leads us to conclude that 3D convolutions are suitable for creating temporally coherent downscaled images (Vondrick et al., 2016; Tran et al., 2015). In combination with linear temporal interpolation within G, 3D convolutions are a crucial factor for the generation of these consistently evolving rain fields. 3D convolutional layers in D may also contribute to spateGANs high temporal consistency, which is supported by a similar application for precipitation nowcasting (Ravuri et al., 2021). However, in our use case their impact on structural precision, that is, the localization of rain cells, might be more significant.

Model Limitations

Despite its potential, 3D convolution has certain limitations and its usefulness for video generation is still a matter of debate (Saito et al., 2017). The main challenge is that the possible amount of exploitable large-scale and long-term spatio-temporal correlations is not arbitrarily expandable. It depends on the model architecture and model depth which define the receptive field size. Furthermore, also the spatial and temporal dimensions of the training samples are important, since model extrapolation capabilities beyond this dimension might be highly limited. Overall, the potential is therefore tied to the available GPU resources, while the memory requirements of 3D convolution are substantial. On the other hand, fully convolutional networks allow for arbitrary input dimensions and we found that spateGANs architecture and depth is sufficient to achieve high performance within the super-resolution downscaling approach. While the model predictions are already spatially and temporally consistent beyond the training sample dimensions it remains unclear if the performance could be further increased by leveraging longer time scales and a larger spatial extent during training. We assume that in the case of downscaling global climate data, an increase in the model's receptive field might be crucial to realize the full potential.

Distribution of Downscaled Rainfall

A main objective of a spatio-temporal downscaling model is the ability to accurately reconstruct the distribution of rainfall at a higher spatial and temporal resolution, which is typically characterized by increased variability and extremes. As expected, the FSS of all models declines towards heavier rainfall, which is harder to model due to its rare occurrence and higher spatio-temporal gradients.

Among the evaluated models, spateGAN stands out as the only model that successfully reconstructed rainfall intensities greater than $5\,mm\,h^{-1}$ or $15\,mm\,h^{-1}$, while maintaining a low BIAS ($< 3.6\,\%$). This is a crucial feature that is not provided by the comparison models. Trilinear interpolation shows the lowest BIAS, however, also the lowest downscaling skill in terms of FSS and RAPSD. The CNN predictions show high skill regarding pixel accuracy metrics, distribution error or downscaling skill for small and moderate rain rates. However, the model is not able to skilfully reconstruct strong precipitation intensities. Furthermore, the model fails to preserve the overall rain sum, maintained within the coarsened input data showing a strong negative BIAS (-22.22 %). We therefore emphasize, as also described in Leinonen et al. (2021), that MAE and KS statistics should be interpreted with caution, as the results could be highly affected by the large amount of small values within the skewed rainfall distribution. They are therefore not suitable to account for the model's ability to recover the target rain distribution containing also extreme values. Furthermore, they can lead to poor metrics, even if models are able to generate rain cells with correct structure and intensity, since these rain cells might be slightly off positioned within the underdetermined downscaling problem and the stochasticity of the solution.

Unexpected Findings

Our analysis of long aggregations (several thousand time steps) of generated rain fields revealed the presence of local biases in the form of recurrent structures. With varying intensity and frequency, they could be observed within the predictions of all ANN models. It is known that GANs can produce artefacts (Karras et al., 2019). However, in our case they were not detectable in single images, e.g., by calculating the power spectrum density. Preliminary results indicate that such model behavior is not unique to the models used in this study, as other prominent ANN downscaling models might also be affected by this behaviour.

While the training images for our models are selected at random locations, reducing the influence of topography, the generated structures are not completely random. Instead, they might follow a spatial or even geometric regularity which is contradictory to the physical principle of emerging rain fields. This does not imply that the downscaling performance of the models is reduced, but can be a seen as a limitation and should be a known feature to be tested. In an effort to minimize the occurrence of these structures, we presented a model with a sophisticated architecture and interpolation technique. Furthermore, we also considered the appearance of these structures in the selection process of the final models (see Section 2.8). Despite this, we were unable to completely eliminate them. Our analysis revealed that a discriminator with many parameters (e.g. G : 2 million., D: 10 million) might lead to an earlier and more intense occurrence of these phenomena. Additionally, we assume that the combination of up and down-sampling layers and their kernel sizes also have an influence. To fully understand the underlying mechanisms responsible for the observed structures, a comprehensive investigation involving the comparison of various hyper-parameterizations would be required. Given the computational effort for training one model, this investigation is beyond the scope of this study and will be left for future research. In the geosciences not only single instances, but also the aggregation of many instances is of importance. Therefore, we emphasize that it is not sufficient to only analyze single predictions, but also the models abilities to fulfill global properties like the climatology of the modeled target variable.

5 Conclusion

Downscaling the output of global climate models is a long-standing problem for providing high-resolution information which is needed to develop adaptation and mitigation strategies in a changing climate. We presented spateGAN, a deep generative model, for simultaneous spatio-temporal downscaling of low-resolution precipitation data. The model was trained using ten years of high-resolution country-wide weather radar rainfall observations in Germany. Our results demonstrated that 3D convolution in combination with conditional generative adversarial networks is an effective tool for leveraging spatio-temporal structures embedded in the low-resolution domain to generate temporally consistent high-resolution rainfall fields and reconstruct the scale dependent extreme value distribution with high skill. This confirms that super-resolution deep learning approaches can be extended to the time dimension to map, in addition to the spatial variability, also the temporal evolution of atmospheric variables.

While a visual inspection leads to the conclusion that generated rain cells look realistic, we found the power spectrum analysis and the Fractions Skill Score to be useful metrics for quantifying this property. Pixel accuracy metrics like the mean absolute error were unable to distinguish between models with high or low skill in generating realistic rain fields. Especially our findings about recurrent structures in downscaled rainfall fields show that a structural analysis is very important in order to mitigate these issues. Overall, the chosen analysis was able to prove that models like spateGAN show great potential to complement and even outperform the capabilities of traditional downscaling methods due to their high performance, computational efficiency and the ability to process arbitrary spatial and temporal input dimensions.

One of the primary purposes of spateGAN is the application for downscaling global climate model outputs. We envision that the approach for this task will have to extend the presented video super-resolution approach, since model outputs are biased with respect to the observed precipitation. Therefore, requirements for the downscaling model would include an additional bias correction step. The potential for bias correction and spatial downscaling of weather forecast data using generative networks has in been demonstrated in L. Harris et al. (2022) and Price and Rasp (2022) and resulted in a performance reduction compared to downscaling coarsened observations. A similar result should be expected for spatio-temporal downscaling. However, we assume that with increased lead time a decoupling of model projections from real observations is the reason for the performance decline and not the insufficient potential of the deep learning approach. Additionally, further studies will have to prove if the generated precipitation fields are suitable, e.g. for simulating the characteristics of flood events under future climate conditions. This work should provide a solid basis for such future studies by not only presenting a high performance downscaling model, but also the analytical framework for a comprehensive analysis of the model performance.

Appendix A Supplementary Figure

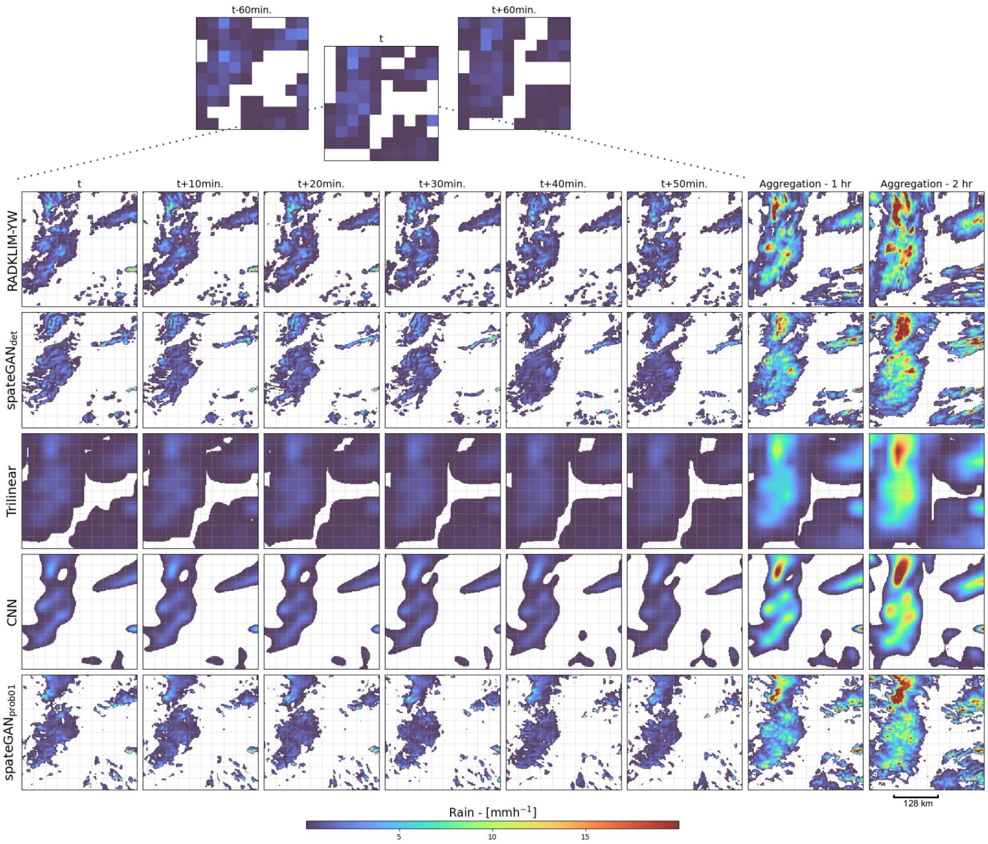

Figure A1. Detailed case study as in Figure 3 for a third event, with a mixture of convective and stratiform rain.

Open Research

The results and models can be reproduced by the publicly available RADKLIM-YW weather radar composite (Winterrath et al., 2018). The CNN and spateGANs were implemented and optimized in a Python framework using TENSORFLOW (version: 2.6) (Developers, 2022). The data and spateGAN models, available in `https://doi.org/10.5281/zenodo.7636929`, provide further insight into the presented spatio-temporal downscaling approach.

Acknowledgments

This study was supported by SCENIC (Storyline Scenarios of Extreme Weather, Climate, and Environmental Events along with their Impacts in a Warmer World) funded by HGF-Innopool, the German Research Foundation (Grant CH 1785/1-2) and the Federal Ministry of Education and Research (Grant 13N14826). We acknowledge support by the KIT-Publication Fund of the Karlsruhe Institute of Technology. Open Access funding enabled and organized by Projekt DEAL.

References

Allan, R. P., & Soden, B. J. (2008). Atmospheric Warming and the Amplification of Precipitation Extremes. *Science*, *321*(5895), 1481–1484. Retrieved 2022-12-20, from https://www.science.org/doi/10.1126/science.1160787 (Publisher: American Association for the Advancement of Science) doi: 10.1126/science.1160787

Baño-Medina, J., Manzanas, R., & Gutiérrez, J. M. (2020). Configuration and intercomparison of deep learning neural models for statistical downscaling. *Geoscientific Model Development*, *13*(4), 2109–2124. Retrieved 2022-09-29, from https://gmd.copernicus.org/articles/13/2109/2020/ (Publisher: Copernicus GmbH) doi: 10.5194/gmd-13-2109-2020

Berg, P., Moseley, C., & Haerter, J. O. (2013). Strong increase in convective precipitation in response to higher temperatures. *Nature Geoscience*, *6*(3), 181–185. Retrieved 2022-12-20, from https://www.nature.com/articles/ngeo1731 (Number: 3 Publisher: Nature Publishing Group) doi: 10.1038/ngeo1731

Chen, D., Rojas, M., Samset, B., Cobb, K., Diongue Niang, A., Edwards, P., ... Tréguier, A.-M. (2021). Framing, Context, and Methods. In V. Masson-Delmotte et al. (Eds.), *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 147–286). Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press. (Type: Book Section) doi: 10.1017/9781009157896.003

Chwala, C., & Polz, J. (2021). *cchwala/radolan_to_netcdf.* Zenodo. Retrieved 2022-12-21, from https://zenodo.org/record/4452204 doi: 10.5281/ZENODO.4452204

Developers, T. (2022). *TensorFlow.* Zenodo. Retrieved 2023-01-18, from https://zenodo.org/record/4724125 doi: 10.5281/ZENODO.4724125

Dong, C., Loy, C. C., He, K., & Tang, X. (2016). Image Super-Resolution Using Deep Convolutional Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(2), 295–307. (Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence) doi: 10.1109/TPAMI.2015.2439281

Gneiting, T., & Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. Retrieved 2022-12-01, from http://www.tandfonline.com/doi/abs/10.1198/016214506000001437 doi: 10.1198/016214506000001437

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative Adversarial Nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (Vol. 27). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

Harris, D., Foufoula-Georgiou, E., Droegemeier, K. K., & Levit, J. J. (2001). Multiscale Statistical Properties of a High-Resolution Precipitation Forecast. *Journal of Hydrometeorology*, *2*(4), 406–418. Retrieved 2022-12-01, from https://journals.ametsoc.org/view/journals/hydr/2/4/1525-7541_2001_002_0406_mspoah_2_0_co_2.xml (Publisher: American Meteorological Society Section: Journal of Hydrometeorology) doi: 10.1175/1525-7541(2001)002⟨0406:MSPOAH⟩2.0.CO;2

Harris, L., McRae, A. T. T., Chantry, M., Dueben, P. D., & Palmer, T. N. (2022). A Generative Deep Learning Approach to Stochastic Downscaling of Precipitation Forecasts. *Journal of Advances in Modeling Earth Systems*, *14*(10), e2022MS003120. Retrieved 2023-02-10, from https://onlinelibrary.wiley.com/doi/abs/10.1029/2022MS003120 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2022MS003120) doi:

10.1029/2022MS003120
IFRC. (2021). *World Disasters Report 2020 \textbar IFRC.* Retrieved 2022-12-17, from https://www.ifrc.org/document/world-disasters-report-2020
Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning* (pp. 448–456). PMLR. Retrieved 2023-02-14, from https://proceedings.mlr.press/v37/ioffe15.html (ISSN: 1938-7228)
Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-Image Translation with Conditional Adversarial Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5967–5976). (ISSN: 1063-6919) doi: 10.1109/CVPR.2017.632
Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer Vision – ECCV 2016* (pp. 694–711). Cham: Springer International Publishing. doi: 10.1007/978-3-319-46475-6_43
Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*(1), 27. Retrieved 2022-12-22, from https://doi.org/10.1186/s40537-019-0192-5 doi: 10.1186/s40537-019-0192-5
Karras, T., Laine, S., & Aila, T. (2019). A Style-Based Generator Architecture for Generative Adversarial Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4396–4405). (ISSN: 2575-7075) doi: 10.1109/CVPR.2019.00453
Kashinath, K., Mustafa, M., Albert, A., Wu, J.-L., Jiang, C., Esmaeilzadeh, S., . . . Prabhat, n. (2021). Physics-informed machine learning: case studies for weather and climate modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *379*(2194), 20200093. Retrieved 2022-02-02, from https://royalsocietypublishing.org/doi/10.1098/rsta.2020.0093 (Publisher: Royal Society) doi: 10.1098/rsta.2020.0093
Kim, J., Lee, J. K., & Lee, K. M. (2016). Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1646–1654). (ISSN: 1063-6919) doi: 10.1109/CVPR.2016.182
Larsen, A. B. L., Sønderby, S. K., Larochelle, H., & Winther, O. (2016). Autoencoding beyond pixels using a learned similarity metric. In *Proceedings of The 33rd International Conference on Machine Learning* (pp. 1558–1566). PMLR. Retrieved 2023-02-14, from https://proceedings.mlr.press/v48/larsen16.html (ISSN: 1938-7228)
Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., . . . Shi, W. (2017). Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 105–114). Honolulu, HI: IEEE. Retrieved 2023-02-14, from http://ieeexplore.ieee.org/document/8099502/ doi: 10.1109/CVPR.2017.19
Leinonen, J., Nerini, D., & Berne, A. (2021). Stochastic Super-Resolution for Downscaling Time-Evolving Atmospheric Fields with a Generative Adversarial Network. *IEEE Transactions on Geoscience and Remote Sensing*, *59*(9), 7211–7223. Retrieved 2022-02-23, from http://arxiv.org/abs/2005.10374 (arXiv: 2005.10374) doi: 10.1109/TGRS.2020.3032790
Lucas, A., Katsaggelos, A. K., Lopez-Tapuia, S., & Molina, R. (2018). Generative Adversarial Networks and Perceptual Losses for Video Super-Resolution. In *2018 25th IEEE International Conference on Image Processing (ICIP)* (pp. 51–55). Athens: IEEE. Retrieved 2023-02-13, from https://

ieeexplore.ieee.org/document/8451714/ doi: 10.1109/ICIP.2018.8451714

Mirza, M., & Osindero, S. (2014). *Conditional Generative Adversarial Nets.* arXiv. Retrieved 2023-02-09, from http://arxiv.org/abs/1411.1784 (arXiv:1411.1784 [cs, stat])

Mu, B., Qin, B., Yuan, S., & Qin, X. (2020). A Climate Downscaling Deep Learning Model considering the Multiscale Spatial Correlations and Chaos of Meteorological Events. *Mathematical Problems in Engineering*, *2020*, e7897824. Retrieved 2021-11-20, from https://www.hindawi.com/journals/mpe/2020/7897824/ (Publisher: Hindawi) doi: 10.1155/2020/7897824

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., ... Anandkumar, A. (2022). *FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators.* arXiv. Retrieved 2023-01-06, from http://arxiv.org/abs/2202.11214 (arXiv:2202.11214 [physics])

Price, I., & Rasp, S. (2022). Increasing the accuracy and resolution of precipitation forecasts using deep generative models. *arXiv:2203.12297 [cs, stat]*. Retrieved 2022-04-06, from http://arxiv.org/abs/2203.12297 (arXiv: 2203.12297)

Pulkkinen, S., Nerini, D., Pérez Hortal, A. A., Velasco-Forero, C., Seed, A., Germann, U., & Foresti, L. (2019). Pysteps: an open-source Python library for probabilistic precipitation nowcasting (v1.0). *Geoscientific Model Development*, *12*(10), 4185–4219. Retrieved 2023-01-18, from https://gmd.copernicus.org/articles/12/4185/2019/ doi: 10.5194/gmd-12-4185-2019

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., ... Mohamed, S. (2021). Skilful precipitation nowcasting using deep generative models of radar. *Nature*, *597*(7878), 672–677. Retrieved 2022-04-20, from https://www.nature.com/articles/s41586-021-03854-z doi: 10.1038/s41586-021-03854-z

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204. Retrieved 2023-02-13, from http://www.nature.com/articles/s41586-019-0912-1 doi: 10.1038/s41586-019-0912-1

Roberts, N. (2008). Assessing the spatial and temporal variation in the skill of precipitation forecasts from an NWP model. *Meteorological Applications*, *15*(1), 163–169. Retrieved 2022-12-01, from https://onlinelibrary.wiley.com/doi/abs/10.1002/met.57 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/met.57) doi: 10.1002/met.57

Roberts, N. M., & Lean, H. W. (2008). Scale-Selective Verification of Rainfall Accumulations from High-Resolution Forecasts of Convective Events. *Monthly Weather Review*, *136*(1), 78–97. Retrieved 2022-12-01, from https://journals.ametsoc.org/view/journals/mwre/136/1/2007mwr2123.1.xml (Publisher: American Meteorological Society Section: Monthly Weather Review) doi: 10.1175/2007MWR2123.1

Saito, M., Matsumoto, E., & Saito, S. (2017). Temporal Generative Adversarial Nets with Singular Value Clipping. In *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 2849–2858). (ISSN: 2380-7504) doi: 10.1109/ICCV.2017.308

Seneviratne, S., Zhang, X., Adnan, M., Badi, W., Dereczynski, C., Di Luca, A., ... Zhou, B. (2021). Weather and Climate Extreme Events in a Changing Climate. In V. Masson-Delmotte et al. (Eds.), *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 1513–1766). Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press. (Type: Book Section) doi: 10.1017/9781009157896.013

Serifi, A., Günther, T., & Ban, N. (2021). Spatio-Temporal Downscaling of Climate

Data Using Convolutional and Error-Predicting Neural Networks. *Frontiers in Climate*, *3*. Retrieved 2022-06-09, from https://www.frontiersin.org/article/10.3389/fclim.2021.656479

Sinclair, S., & Pegram, G. G. S. (2005). Empirical Mode Decomposition in 2-D space and time: a tool for space-time rainfall analysis and nowcasting. *Hydrology and Earth System Sciences*, 11.

Sun, A. Y., & Tang, G. (2020). Downscaling Satellite and Reanalysis Precipitation Products Using Attention-Based Deep Convolutional Neural Nets. *Frontiers in Water*, *2*. Retrieved 2022-03-09, from https://www.frontiersin.org/article/10.3389/frwa.2020.536743

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2015). Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1–9). (ISSN: 1063-6919) doi: 10.1109/CVPR.2015.7298594

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning Spatiotemporal Features with 3D Convolutional Networks. In *2015 IEEE International Conference on Computer Vision (ICCV)* (pp. 4489–4497). (ISSN: 2380-7504) doi: 10.1109/ICCV.2015.510

Vandal, T., Kodra, E., Ganguly, S., Michaelis, A., Nemani, R., & Ganguly, A. R. (2017). DeepSD: Generating High Resolution Climate Change Projections through Single Image Super-Resolution. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1663–1672). Halifax NS Canada: ACM. Retrieved 2022-02-01, from https://dl.acm.org/doi/10.1145/3097983.3098004 doi: 10.1145/3097983.3098004

Vaughan, A., Tebbutt, W., Hosking, J. S., & Turner, R. E. (2022). Convolutional conditional neural processes for local climate downscaling. *Geoscientific Model Development*, *15*(1), 251–268. Retrieved 2022-02-15, from https://gmd.copernicus.org/articles/15/251/2022/ (Publisher: Copernicus GmbH) doi: 10.5194/gmd-15-251-2022

Vondrick, C., Pirsiavash, H., & Torralba, A. (2016). Generating Videos with Scene Dynamics..

Wang, F., Tian, D., Lowe, L., Kalin, L., & Lehrter, J. (2021). Deep Learning for Daily Precipitation and Temperature Downscaling. *Water Resources Research*, *57*(4), e2020WR029308. Retrieved 2021-11-20, from https://onlinelibrary.wiley.com/doi/abs/10.1029/2020WR029308 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2020WR029308) doi: 10.1029/2020WR029308

Wang, X., Lucas, A., Lopez-Tapia, S., Wu, X., Molina, R., & Katsaggelos, A. K. (2019). Spatially Adaptive Losses for Video Super-resolution with GANs. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1697–1701). (ISSN: 2379-190X) doi: 10.1109/ICASSP.2019.8682742

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., ... Loy, C. C. (2019). ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In L. Leal-Taixé & S. Roth (Eds.), *Computer Vision – ECCV 2018 Workshops* (pp. 63–79). Cham: Springer International Publishing. doi: 10.1007/978-3-030-11021-5_5

Winterrath, T., Brendel, C., Hafer, M., Junghänel, T., Klameth, A., Lengfeld, K., ... Becker, A. (2018). *Radar climatology (RADKLIM) version 2017.002; gridded precipitation data for Germany: Radar-based quasi gauge-adjusted five-minute precipitation rate (YW).* Deutscher Wetterdienst (DWD). Retrieved 2022-12-21, from https://opendata.dwd.de/climate_environment/CDC/help/landing_pages/doi_landingpage_RADKLIM_RW_V2017.002-en.html (Artwork Size: approx. 500 MByte per gzip compressed ascii or binary archive with monthly data Medium: gzip compressed and tar packed ascii and binary Pages: approx. 500 MByte per gzip compressed ascii or binary archive with monthly data Version Number: 1 Type: dataset) doi: 10.5676/DWD/RADKLIM_YW_V2017.002

Winterrath, T., Brendel, C., Hafer, M., Junghänel, T., Klameth, A., Walawender, E., … Becker, A. (2017). *Erstellung einer radargestützten Niederschlagsklimatologie* (Tech. Rep. No. 251). Deutscher Wetterdienst. Retrieved 2022-12-21, from https://www.dwd.de/DE/leistungen/pbfb_verlag_berichte/pdf_einzelbaende/251_pdf.pdf?__blob=publicationFile&v=2