RainScaleGAN: A Conditional Generative Adversarial Network for Rainfall Downscaling

MARCELLO IOTTI,[a,b] PAOLO DAVINI,[c] JOST VON HARDENBERG,[c,d] AND GIUSEPPE ZAPPA[e]

[a] *Dipartimento di Fisica e Astronomia (DIFA), Alma Mater Studiorum - Università di Bologna, Bologna, Italy*
[b] *Istituto di Matematica Applicata e Tecnologie Informatiche, Consiglio Nazionale delle Ricerche (CNR-IMATI), Genoa, Italy*
[c] *Istituto di Scienze dell'Atmosfera e del Clima, Consiglio Nazionale delle Ricerche (CNR-ISAC), Turin, Italy*
[d] *Dipartimento di Ingegneria dell'Ambiente, del Territorio e delle Infrastrutture (DIATI), Politecnico di Torino, Turin, Italy*
[e] *Istituto di Scienze dell'Atmosfera e del Clima, Consiglio Nazionale delle Ricerche (CNR-ISAC), Bologna, Italy*

(Manuscript received 22 August 2024, in final form 24 February 2025, accepted 3 April 2025)

ABSTRACT: To this day, accurately simulating local-scale precipitation and reliably reproducing its distribution remains a challenging task. The limited horizontal resolution of global climate models is among the primary factors undermining their skill in this context. The physical mechanisms driving the onset and development of precipitation, especially in extreme events, operate at spatiotemporal scales smaller than those numerically resolved, thus struggling to be captured accurately. To circumvent this limitation, several downscaling approaches have been developed over the last decades to address the discrepancy between the spatial resolution of models' output and the resolution required by local-scale applications. In this paper, we introduce RainScaleGAN, a conditional deep convolutional Generative Adversarial Network (GAN) for precipitation downscaling. GANs have been effectively used in image superresolution, an approach highly relevant for downscaling tasks. RainScaleGAN's capabilities are tested in a *perfect-model* setup, where the spatial resolution of a precipitation dataset is artificially degraded from $0.25° \times 0.25°$ to $2° \times 2°$, and RainScaleGAN is used to restore it. The developed model outperforms one of the leading precipitation downscaling methods found in the literature. RainScaleGAN not only generates a synthetic dataset featuring plausible high-resolution spatial patterns and intensities but also produces a precipitation distribution with statistics closely mirroring those of the ground-truth dataset. Given that RainScaleGAN's approach is agnostic with respect to the underlying physics, the method has the potential to be applied to other physical variables such as surface winds or temperature.

SIGNIFICANCE STATEMENT: Accurately predicting local precipitation is difficult due to the limitations of current climate models. These models struggle to capture the small-scale processes causing precipitation, especially those leading to extreme events. To address this, we developed a new tool that uses advanced artificial intelligence techniques to improve rainfall predictions. This tool takes low-resolution precipitation data and enhances it to high resolution, providing more detailed rainfall patterns. Our results show that the tool performs better than one of the leading existing methods. This advancement could lead to more precise climate projections, better preparation for extreme weather, and suggests further exploration on additional weather variables.

KEYWORDS: Downscaling; Deep learning; Machine learning; Neural networks; Climate models; Artificial Intelligence

1. Introduction

Global climate models (GCMs) are nowadays the primary tools for the investigation of the climate system, its mechanisms, and its changes. Despite the remarkable skill achieved in the recent decades across a wide range of applications, and their continuous evolution, they still lack accuracy in the reproduction of the precipitation distribution (cf. Sha et al. 2020). The most relevant reason for such limitation can be traced back to the spatial resolution at which most of the current state-of-the-art GCMs are run, usually falling within the range of 50–200 km. Such horizontal resolution, much coarser than the typical spatial scale of precipitation and convective structures, can reasonably capture the synoptic and part of the mesoscale atmospheric circulation but is too coarse to accurately represent small-scale phenomena, particularly where proper modeling requires a precise representation of surface meteorological variables on topographically complex terrain. Weather and climate models rely on specific *parameterizations* to tackle this inadequacy, and despite notable improvements in recent years, this approach still presents limitations. In particular, numerical models suffer from an imperfect physical representation of precipitation: they usually simulate convective and stratiform precipitation independently, resulting in an inaccurate precipitation distribution, where typically the occurrence of light rain (drizzle) is overestimated, while dry days and high to extreme events are underestimated (see, e.g., Piani et al. 2010).

Beyond simulating the atmosphere for research purposes, atmospheric modeling serves societally relevant goals. It plays a crucial role in supporting a range of applications, including hydrological modeling, water management, and agriculture. In broader terms, it enables scientific evidence–based decision-making processes for policymakers, engineers, and planners, who need to understand and formulate a response to predicted events. All these applications are highly sensitive to the precipitation input they receive, to its resolution, and to the details

Corresponding author: Marcello Iotti, marcello.iotti@ge.imati.cnr.it

DOI: 10.1175/AIES-D-24-0074.1 e240074

of its fine-scale distribution, thus requiring greater accuracy and a finer spatial distribution than what GCMs provide.

In atmospheric sciences, the term *downscaling* refers to any operation aimed at inferring high-resolution variables from low-resolution data. Many techniques exist, founded on the assumption that the large-scale configuration of the atmosphere strongly influences variables at the local scale. Downscaling techniques can be classified into two main groups, each approaching the task differently (for an in-depth review, see Maraun et al. 2010). On the one hand, *dynamical downscaling* uses high-resolution *regional climate models* nested within low-resolution GCMs, which provide boundary conditions to them (Feser et al. 2011; Rummukainen 2010). While these models have a strong physical basis, they come with large computational costs, limiting their coverage to specific areas and a restricted number of simulations. On the other hand, *statistical downscaling* is a postprocessing technique that establishes statistical relationships between large-scale predictors and small-scale predictands (Wilby and Wigley 1997; Rummukainen 1997; Wilby et al. 1999; Dibike and Coulibaly 2005). Methods within this category are computationally less expensive, yet their calibration relies on high-quality local-scale data, which may not always be available everywhere. Moreover, they might not be easily transferable to different regions of the globe. A particular category of statistical methods is represented by *stochastic downscaling* methods, a form of weather generators (Maraun et al. 2010), which, starting only from large-scale precipitation fields, can generate fine-scale downscaled fields with a realistic spatial correlation structure and amplitude distribution (see, for example, Ferraris et al. 2003).

In recent years, applying machine learning (ML) techniques originally developed in image processing to downscaling tasks has produced remarkable results. The exploration of such techniques in a context different from their origin has been driven by the similarity between downscaling and the so-called image superresolution (upsampling),[1] which is the process of enhancing the resolution of an image (Reichstein et al. 2019). To date, the most successful ML models in the field of image processing are based on convolutional neural networks (CNNs), leading many authors to tackle the downscaling problem using CNNs (Sha et al. 2020; Kumar et al. 2021; Wang et al. 2021). Further advancements have come from applying generative adversarial networks (GANs) (Goodfellow et al. 2014, 2020) to the downscaling problem. The goal of GANs is to train a neural network, called *generator*, to generate examples that mimic the probability distribution of the training data. A complementary neural network, the *discriminator*, is designed to assess the generated examples, distinguishing them from the training data and then encouraging the generator to enhance its performance. Ledig et al. (2017) applied GANs to image superresolution, while Leinonen et al. (2021) introduced a recurrent superresolution GAN able to generate ensembles of plausible high-resolution atmospheric fields from their low-resolution (upscaled) counterparts. Ravuri et al. (2021) addressed the problem of the so-called *nowcasting*, developing a deep generative model for the probabilistic short-term prediction of radar-measured precipitation. Harris et al. (2022) and Price and Rasp (2022) extended the problem addressed by Leinonen, building models mapping from multiple low-resolution atmospheric fields (including precipitation) from a numerical weather prediction model to high-resolution radar-measured precipitation. More recently, Annau et al. (2023) developed a superresolution GAN-based model trained on nonidealized pairs consisting of low-resolution (80 km) reanalysis data and 10-m wind component fields from a convection-permitting (4 km) model driven by the same low-resolution dataset. Their model aims to reproduce fine-scale details consistent with the convection-permitting simulation, effectively capturing its internal variability.

In this paper, we will demonstrate how a GAN with a simple architecture can effectively downscale precipitation. By relying solely on the low-resolution precipitation field as a predictor, our approach can be easily generalized to any part of the globe, as it is independent of explicitly incorporating the topographic features of the geographical region under investigation, nor does it require additional external sources of information. However, it is important to note that extending the method to other regions would require additional training with region-specific precipitation data. We will conduct the training and testing of the model in the so-called *perfect-model setup* (pure superresolution, cf. Harris et al. 2022), reducing the resolution of training data through spatial aggregation, and using our model to restore the lost original resolution. We will demonstrate how the generated dataset closely mirrors the statistical properties of the original dataset. Additionally, our trained generator proves to be more effective in producing high-resolution precipitation fields compared to RainFARM (Rebora et al. 2006; D'Onofrio et al. 2014; Terzago et al. 2018), a state-of-the art stochastic downscaling method.

The structure of the paper is as follows: Section 2 presents the data used in our experiments and their preprocessing. In section 3, we define the task we addressed, describe the model architecture, outline the training process, list the metrics used to assess the performance of the model, and briefly introduce RainFARM, the alternative downscaling method used as a baseline to assess RainScaleGAN's skills. Section 4 presents the results of the experiments, describing the training process, model validation and testing, with the final comparison with RainFARM. The final section 5 discusses the results, the limitations of the adopted framework, and possible future developments.

2. Data

The ERA5 (Hersbach et al. 2023) reanalysis data for total precipitation have been used throughout the entire machine learning exercise and for conducting the downscaling process using RainFARM. This variable represents the cumulative amount of liquid and solid precipitation, resulting from both large-scale and convective precipitation. The spatial covering

[1] Note that, somewhat confusingly, the term *upsampling* (*downsampling*) in the field of image processing refers to the process of increasing (decreasing) the resolution of an image. This is exactly the opposite of the terms *upscaling*/*downscaling* used in meteorology.

is global, with a resolution of 0.25° × 0.25°, and the temporal resolution equals 1 h.

To demonstrate our model, we chose a region of interest centered on the Alpine arch, spanning latitudes 38°–53.75°N and longitudes 3°–18.75°E. This region includes both sea and land for the majority of the Italian Peninsula, Austria, the Czech Republic, the central and southern parts of Germany, Switzerland, the Netherlands, Belgium, the eastern part of France, and some portions of the neighboring countries. Such a choice is rather arbitrary, but our model is designed to be agnostic regarding the region to which it is applied. Additionally, the selected region encompasses a topographically complex terrain, due to the presence of orography, making it a suitable testbed for a rainfall downscaling technique.

a. Data source and preprocessing

The ERA5 total precipitation has been obtained through the Copernicus Climate Data Store (CDS) [Copernicus Climate Change Service (C3S) 2023]. In this archive, ERA5 data are interpolated onto a regular latitude–longitude grid. The total precipitation is derived from short (18-h) forecasts, run twice a day from the 0600 and 1800 UTC analyses. The accumulation is carried out for the hour ending at the date and time of validity. We computed the daily precipitation by taking the average of the hourly values within the same date and then multiplying by 24 (the number of hours in a day). This value is not coincident with the precipitation actually accumulated during the corresponding 24 h, as the precipitation with valid time 0000 UTC is accumulated from 2300 to 2359 UTC of the previous day. In the present study, we do not plan a comparison with measured data; therefore, such inconsistency is not relevant.

The data related to the domain of interest have been extracted, without performing any further spatial interpolation, resulting in precipitation fields of 64 × 64 grid points (a box of approximately 1800 km × 1300 km). A spatial filter has been applied to the dataset to exclude days with extremely low precipitation, as a measure to counterbalance part of the drizzle problem. Additionally, we observed that exposing the GAN to nonmeaningful samples slows down its convergence. The filter is implemented by computing the spatial average of the precipitation across the entire domain for each sample of the dataset. Days are excluded from the dataset if this quantity falls below a small threshold, arbitrarily set at 1 mm. The resulting two-dimensional daily precipitation fields constitute the *examples* used for training the GAN.

Feature scaling is a fundamental preprocessing step in most ML tasks. It enhances the convergence speed and performance of ML optimization algorithms, effectively preventing gradient descent issues. Moreover, it helps in handling skewed data and reducing the impact of outliers, balancing the influence of features. Therefore, the following transformations are applied to the original precipitation rate x (mm day^{-1}):

1) Square root transformation $\sqrt{x}$.
2) Rescaling (min–max normalization), applied *separately* to each grid point according to the following formula:

$$x_{\text{scaled}} = m + \frac{x - x_{\min}}{x_{\max} - x_{\min}}(M - m), \tag{1}$$

where $x_{\min}$ and $x_{\max}$ are the minimum and the maximum values of the time series for that grid point, and the chosen feature range is $[m, M] = [-1, 1]$.

As precipitation has a strongly positively skewed distribution, the reexpression (see, e.g., Wilks 2011) using the square root transformation contributes to obtaining a more symmetric distribution, facilitating the analysis of data and improving the performance of the ML model. Additionally, it avoids the issue of zeros related to the commonly used logarithmic transformations. Regarding the latter transformation (min–max scaling), since the value of precipitation over each grid point can be considered a feature influenced by the underlying topography and specific precipitation-generating processes, the rescaling ensures a consistent treatment of the entire precipitation field under consideration.

b. Data subsets

Data selection is critical in constructing a data-driven model. Proper sampling of predictor variability is essential for achieving model generalizability. In ML practice, it is standard to divide the data into training, validation, and test subsets, ensuring that these subsets are independent of each other.

However, meteorological data can be conceptualized as time series that are intrinsically autocorrelated over finite spatial and temporal domains. Therefore, the standard procedure often used in deep learning research—extracting random samples from the available data, assuming each instance is independent, and arbitrarily assigning them to the training, validation, and test sets—is inappropriate. This approach can overestimate the skill of the model because random sampling introduces correlations among the three subsets, thereby incorporating information into the test set that has already been used in training. To address this, we adopt the strategy of random block sampling (Schultz et al. 2021), where the dataset is split into blocks with durations much greater than the period of time considered to contribute the most to autocorrelation (a few days). A downside of this method is that it assumes there are no significant long-term trends in the distribution of precipitation, which contrasts with the effects of climate change. Nevertheless, it can serve as a useful indicator of the effectiveness of the ML model in handling such statistical changes over time, assessing its robustness and applicability, for example, in the context of climate projections.

The ERA5 dataset we downloaded, spanning the years from 1940 to 2022, has been divided into three consecutive subsets: data from 1940 to 1998 are used as the training set (15692 examples after filtering), data from 1999 to 2010 are used as the validation set (3261 examples), and data from 2011 to 2022 are used as the test set (3134 examples). To apply minibatch stochastic gradient descent (Goodfellow et al. 2016), shuffling is applied to each of these three subsets. This shuffling aims to ensure independence between the examples within each minibatch, as well as between the minibatches themselves.

3. Methods

a. Definition of the task

An inherent challenge in evaluating a new downscaling technique is the unavoidable presence of biases between the

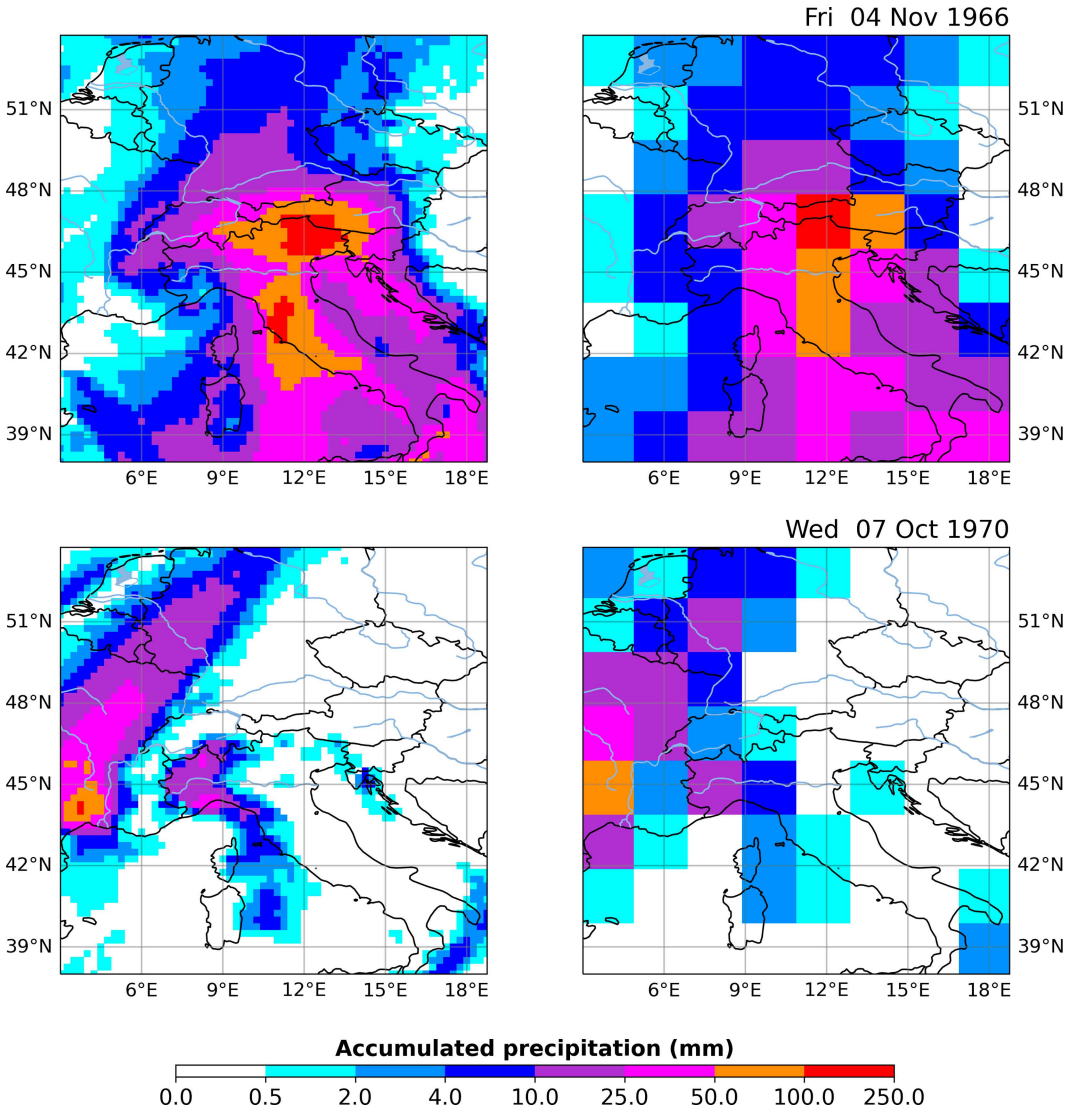

FIG. 1. (left) the ERA5 daily total accumulated precipitation and (right) the corresponding coarsened version for two sample days (4 Nov 1966 and 7 Oct 1970). The original ERA5 examples have a spatial resolution of 0.25° × 0.25° and consist of 64 × 64 grid points. The coarsened versions, obtained with an upscaling factor of 8, have a spatial resolution of 2° × 2° and consist of 8 × 8 grid points.

predictor and target datasets. To circumvent this issue, we adopt what in the literature is referred to as a *perfect-model* setup (Terzago et al. 2018). This involves taking a high-resolution precipitation dataset, measured or simulated, and artificially degrading its spatial resolution through an aggregation operation. The downscaling model is then tasked with restoring the lost resolution based on this smoother, low-resolution precipitation field. This allows for the assessment of the skill of the downscaling method, measuring whether the produced field reflects the correct rainfall patterns and statistical properties of the true field (Rebora et al. 2006). In terms closer to those used in the machine learning domain, it is also called *pure superresolution* (Harris et al. 2022).

To apply this procedure to our study, we performed a spatial aggregation (upscaling) on the ERA5 daily total precipitation, reducing its spatial resolution by a factor of 8, from 0.25° × 0.25° to 2° × 2°. The operation consists in taking the average of precipitation across groups of 8 × 8 adjacent grid cells. The resulting coarsened field covers an area of 8 × 8 grid cells. Figure 1 displays some examples of the outcome of the described operation.

b. *Model architecture*

The model we constructed is a conditional GAN (Mirza and Osindero 2014), i.e., a GAN in which both the generator and discriminator receive, as additional input, conditioning data aimed at directing the generation process. In our case, these conditioning data consist of the low-resolution version of the daily precipitation field to be downscaled. This source of information is conditioning in the sense that it provides the

large-scale structure of the precipitation field that the generated fine-scale example is required to adhere to. The task of the generator is to produce a precipitation field at the target spatial resolution, using an input composed of the following:

- The corresponding low-resolution conditioning field.
- A source of noise, in this case an array of random numbers drawn from a normal distribution with mean 0 and standard deviation 0.02.

The noise source in the generator implies that it is capable of producing an indefinite number of examples consistent with the structure of the low-resolution conditioning field. The task of the discriminator is to distinguish the predictions of the generator from the corresponding "ground-truth" fields from the training set. The discriminator is fed with either ground-truth or generated examples, each one together with the corresponding low-resolution precipitation field, always drawn from the upscaled version of the training data. Please note that the upscaled counterparts of the generated precipitation are never used. Due to the architecture of the neural networks we implemented for the two components of RainScaleGAN, inputs must be concatenated. In the case of the discriminator, this operation requires that input fields have the same spatial dimensions. We thus performed a nearest-neighbor remapping on the low-resolution dataset, generating a rainfall field with the same information content, but with spatial resolution matching the one of the target, high-resolution dataset. The output of the discriminator is used to calculate the loss function for both the discriminator itself and the generator. This way, the discriminator guides the training process, providing a feedback to the generator, ideally enabling it to improve its performance during the training process. Figure 2 presents an overview of the described process, illustrating the interplay between information sources, models, and their outputs during the training phase.

Figure 3 illustrates the architecture of the generator and discriminator, both implemented as deep convolutional artificial neural networks. The input to the generator is a one-dimensional array constructed by flattening the low-resolution input field to be downscaled and concatenating it with the array of random numbers mentioned above. This input is initially mapped through a dense layer to a three-dimensional tensor of appropriate size, depending on the number of grid points in the precipitation field to be generated. Following, there are four blocks of layers designed to map this tensor to the target field. Such number of blocks depends on the spatial extent of the target rainfall field and, ultimately, on the upscaling factor, i.e., the ratio between the spatial resolutions of the low- and high-resolution fields. The essential components of these blocks consist of a two-dimensional upsampling layer followed by a two-dimensional convolutional layer, with increasing spatial dimensions (height and width) and decreasing number of filters. Specifically, the upsampling layers have upsampling factors of (2, 2), doubling the height and width of the tensor they receive, while the convolutional layers have a stride of 1 in both directions and number of filters equal to 256, 128, 64, and 32, respectively. Batch normalization (Ioffe and Szegedy 2015) and

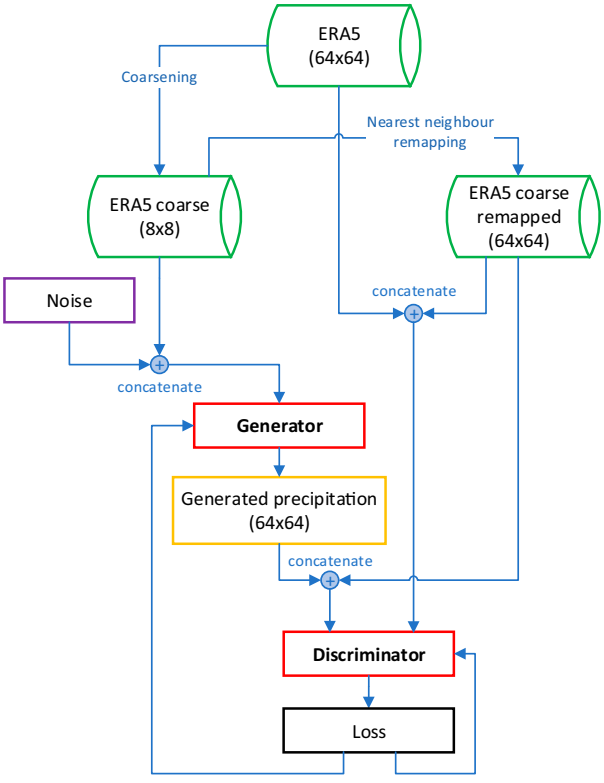

FIG. 2. Information flow during the model training.

leaky rectified linear unit (ReLU) activation with a negative slope of 0.2 are applied at the end of both the input block and each of the aforementioned convolutional blocks. The network concludes with a two-dimensional convolutional layer with a single filter, employing the hyperbolic tangent as activation. This layer aims to generate the final image corresponding to the rainfall field at the target resolution.

The structure of the discriminator mirrors that of the generator. It takes as input pairs of ground-truth/generated high-resolution rainfall fields and their corresponding low-resolution rainfall fields which—as mentioned above—have been remapped with the nearest-neighbor method. These fields are concatenated along their depth (i.e., the images are stacked) and then passed to the discriminator. Four blocks consisting of convolutional layers with strides (2, 2), to reduce the height and width of the input by half at each layer, and increasing number of filters (64, 128, 256, 512) are employed. Each block is activated using the leaky ReLU function with a negative slope of 0.2. The network concludes with a single-filter convolutional layer, densely connected to a single unit, with linear activation (i.e., no final activation is applied).

The size of the convolutional filters, in both the generator and the discriminator, is a parameter that we optimized during the validation phase. The optimal generator and discriminator that we selected have approximately 4.3 and 2.8 million trainable parameters, respectively (compare with the training set size: 15692 examples, each consisting of a 64×64 grid-point precipitation field).

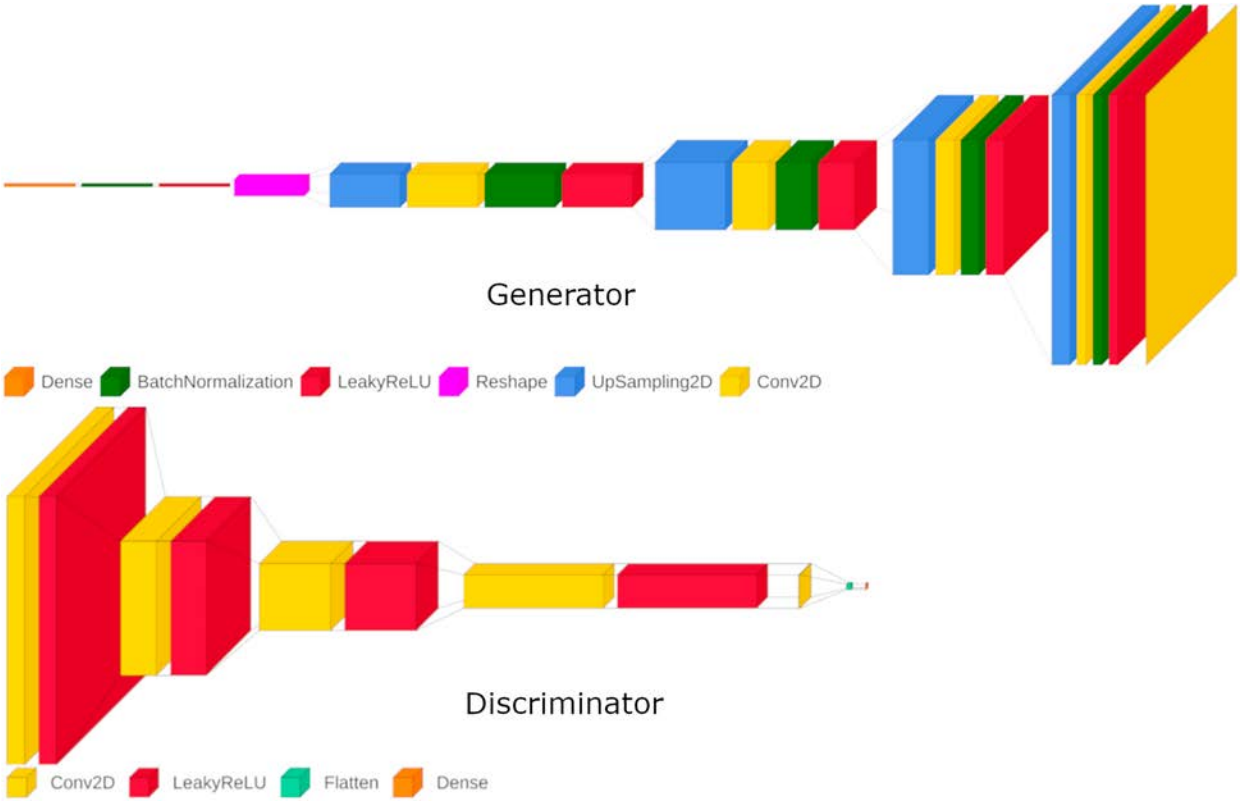

FIG. 3. The architecture of the networks composing RainScaleGAN. The input layers of the two networks (the coarse image together with a noise source for the generator, the generated/ground-truth image together with the corresponding coarse image for the discriminator) are not shown. (top) Generator architecture. The upsampling layers have upsampling factors of (2, 2), thereby doubling the number of rows and columns of their input. The intermediate convolutional layers have a number of kernels equal to 256, 128, 64, and 32, respectively. The final convolutional layer has a single kernel and is activated with the hyperbolic tangent function. (bottom) Discriminator architecture. The convolutional layers, except the last one, have a number of kernels equal to 64, 128, 256, and 512, respectively, and strides (2, 2). Each of them halves the height and width of the field it receives as input.

c. Training

Training a GAN involves the simultaneous training of two models: the discriminator D, which aims to maximize the probability that it assigns the correct label to real (from the training set) and fake (from the generator) examples, and the generator G, which aims to maximize the probability that D mistakenly assigns the label "real" to generated examples. In other words, the training is a minimax game with a value function $V(G, D)$ (Goodfellow et al. 2014):

$$\min_G \max_D V(D, G) = \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}[\log(D(\mathbf{x}))] + +\mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))], \quad (2)$$

where p_{data} and $p_{\mathbf{z}}$ are the distributions of the training data $\mathbf{x}$ and of a noise variable $\mathbf{z}$, respectively. Under appropriate assumptions (cf. Goodfellow et al. 2014), the minimax game expressed by Eq. (2) translates into minimizing the Jensen–Shannon divergence[2] between the distribution of training data and that of generated data. This divergence may not be continuous with respect to the generator parameters (Arjovsky et al. 2017), and its minimization often leads to discriminator saturation with resulting vanishing gradients (Gulrajani et al. 2017). Wasserstein GANs (WGANs) (Arjovsky et al. 2017) are designed to address these issues. The training objective of a WGAN is expressed by

$$\min_G \max_{D\in\mathcal{D}} \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[D(\mathbf{x})] - \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[D(G(\mathbf{z}))], \quad (3)$$

where $\mathcal{D}$ is the set of 1-Lipschitz functions. Once the discriminator (referred to as the *critic* in the foundational paper) is optimized, the minimization of the previous value function with respect to the generator parameters minimizes the Earth-Mover (Wasserstein-1) distance between the distributions of the training and generated data.

The value function based on this distance exhibits better properties than the original value function, making the optimization of the generator simpler. Following Gulrajani et al. (2017), we enforce the Lipschitz constraint on D by introducing a gradient penalty term in the discriminator loss.

[2] The Jensen–Shannon divergence is a measure of the similarity between two probability distributions.

The GAN framework extends to a conditional model by incorporating auxiliary information **y** into both the generator and discriminator (Mirza and Osindero 2014). In a downscaling task, this auxiliary information consists of the low-resolution field to be refined. Thus, the objective functions for the generator and discriminator of the conditional WGAN with gradient penalty (WGAN-GP) used in this study are

$$\begin{aligned} L_D = {} & \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[D(\mathbf{x}|\mathbf{y})] \\ & + -\mathbb{E}_{\mathbf{z}\sim p_z(\mathbf{z})}[D(G(\mathbf{z}|\mathbf{y})|\mathbf{y})] \\ & + + \lambda(\mathbb{E}_{\hat{\mathbf{x}}\sim \hat{P}_{\mathbf{x}}}[(\|\nabla_{\hat{\mathbf{x}}} D(\hat{\mathbf{x}})\|_2 - 1)^2]), \end{aligned} \tag{4}$$

$$L_G = \mathbb{E}_{\mathbf{z}\sim p_z(\mathbf{z})}[D(G(\mathbf{z}|\mathbf{y})|\mathbf{y})], \tag{5}$$

with λ being a constant representing the weight of the gradient penalty term, set equal to 10 in accordance with Gulrajani et al. (2017), and the samples $\hat{\mathbf{x}}$ are defined by

$$\hat{\mathbf{x}} = \boldsymbol{\epsilon}\mathbf{x} + (1 - \boldsymbol{\epsilon})G(\mathbf{z}), \tag{6}$$

where $\boldsymbol{\epsilon}$ is a random number drawn from a uniform distribution $U[0, 1]$. Both the generator and the discriminator aim to maximize their respective objective functions as defined above. Unlike some other conditional GAN approaches for downscaling, we did not implement a content loss term. The relevance of content loss is evident in studies that involve observations (e.g., Harris et al. 2022) or aim to emulate specific characteristics of the process generating the target dataset (e.g., Annau et al. 2023), often with a stronger emphasis on improving per-gridcell metrics. In contrast, its role appears to be less prominent in perfect-model setups like ours (cf. Leinonen et al. 2021). Additionally, this choice allows us to assess the impact of the objective function of the generator, as defined in Eq. (5)—commonly referred to as the *adversarial component*—on effectively guiding the superresolution task as formulated in this study, independent of any content loss.

Following Arjovsky et al. (2017), during the training cycle of RainScaleGAN, we alternate between five iterations of training for the discriminator and one iteration of training for the generator. The Adam optimizer (Kingma and Ba 2017) with a learning rate of 2×10^{-4} has been chosen for both the neural networks.

d. Skill metrics

To assess the performance of the model, as well as to monitor training and conduct validation, we employed the following set of metrics.

As a simple indicator of the quality of the generated precipitation fields, useful for evaluating the convergence of the training process, we use the root-mean-square error:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_{\text{true},i} - x_{\text{gen},i})^2}, \tag{7}$$

where i indexes the grid points and N is their total number. To have a more precise evidence of the ability of the GAN to reconstruct the spatial structure and variability of the generated rainfall field, we calculate the log-spectral distance (LSD) between the spatial radial spectrum of the generated dataset and the corresponding spectrum of the ERA5 dataset:

$$\text{LSD} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(10\log_{10}\frac{P_{\text{true},k}}{P_{\text{gen},k}}\right)^2}. \tag{8}$$

The power spectra P_{true} and P_{gen} are obtained by performing the Fourier transform in the physical (two-dimensional) space of the precipitation field, averaging along the time axis over all the examples within the dataset in question. Then, a binned average is applied in the k space over K bins, each bin being centered on one of the discrete Fourier wavenumbers that can be defined in the physical space. This operation is equivalent to collapsing over all angular directions of the two-dimensional spectrum, obtaining a one-dimensional spectrum (cf. Harris et al. 2022).

The two metrics presented above (henceforth collectively referred to as *image metrics*) are borrowed from the practice of image processing. Although they constitute an important reference point for evaluating the model's skill, they lack in describing the statistical, and in some sense physical, properties of the generated dataset. To provide a more comprehensive assessment of these properties, considering their importance especially within the context of climate studies, we extended our suite of metrics to include a set of basic statistics for the generated dataset. These statistics were not only monitored during training but were also crucial for model selection and in the final assessment of the trained model, facilitating a comparison with the alternative downscaling method. In the following, we will collectively refer to these metrics as *statistical metrics*. Specifically, we considered the climatology and the standard deviation of the time series for daily total precipitation, calculated grid-point-wise:

$$\text{Clim}(i, j) = \frac{\sum_t x_t^{(i,j)}}{T}, \tag{9}$$

$$\text{SD}(i, j) = \sqrt{\frac{\sum_t [x_t^{(i,j)} - \text{Clim}(i,j)]^2}{T}}, \tag{10}$$

where $x_t^{(i,j)}$ is the amount of the daily total accumulated precipitation on the grid point (i, j) for the day t, while T is the total number of days in the dataset. Moreover, we compute the 95th and 99th percentiles of daily total precipitation, once again on a grid-point-wise basis. The climatology and standard deviation enable the assessment of the mean statistical properties of the generated dataset. The uppermost percentiles are important in evaluating the GAN's capability to accurately capture both the magnitude and the localization of extreme events.

e. Alternative method

As a baseline for comparing the performance of the constructed model, we chose RainFARM (Rebora et al. 2006), a well-established method for rainfall downscaling that relies on a nonlinear transformation of a Gaussian random field. A detailed description of the RainFARM approach, as well as its subsequent refinement by Terzago et al. (2018), is provided

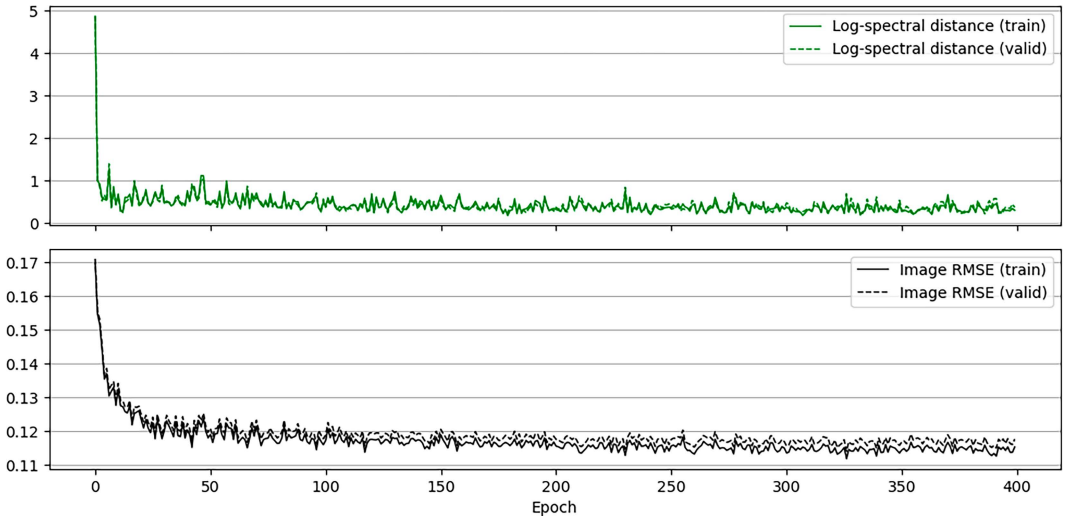

FIG. 4. Evolution of image metrics throughout the GAN training process. (top) LSD between the spatial radial spectra of the generated precipitation and that of the corresponding true dataset. (bottom) RMSEs between the generated images and their corresponding true counterparts. The solid lines refer to the training dataset (1940–98), while the dashed lines correspond to the validation dataset (1999–2010).

in the appendix. In this work, we used this latest version of RainFARM when comparing with RainScaleGAN.

To ensure consistency between the data provided to the GAN during training, RainFARM was run using the slope of the power spectrum of the upscaled ERA5 training set (1940–98). The climatology of these data at the original resolution was used to compute the corrective weights. Once these parameters were determined, RainFARM was applied to the upscaled ($2° \times 2°$) ERA5 test set to generate precipitation fields at the target resolution ($0.25° \times 0.25°$). These fields were then used as a baseline for evaluating the GAN.

4. Results

a. *Training analysis*

During training, we expect the generated dataset to progressively become similar to the training dataset. This implies that its statistical properties will converge to those of the training dataset. Unlike a standard GAN, RainScaleGAN—which is based on a Wasserstein GAN—has loss functions correlated with both the convergence of the generator and the quality of the generated data (cf. Arjovsky et al. 2017; Gulrajani et al. 2017). However, to have a more meaningful perspective on the quality of the climate delivered, we decided to rely on the set of metrics defined in section 3d to monitor the training process.

Figure 4 shows the evolution of the log-spectral distance and of the root-mean-square error between the (spectra of) generated precipitation fields and their corresponding true counterparts, throughout RainScaleGAN's training process. Similarly, Fig. 5 shows the evolution of the root-mean-square errors for climatology, standard deviation, and the 95th and 99th percentiles, with respect to the corresponding quantities for the true datasets. In detail, the calculation of these metrics proceeded as follows:

- At the end of each training epoch, the generator, with parameters fixed at the last update, was used to reconstruct both the training set (1940–98) and the validation set (1999–2010), from the corresponding ERA5 upscaled versions.
- We computed the root-mean-square errors between the generated datasets and their ground-truth ERA5 counterparts, for both the training and the validation sets. The data used in this computation are bounded within the range $[-1, 1]$, originating from a dataset subjected to the preprocessing operations outlined in section 2a.
- The generated train and validation datasets were denormalized, applying the inverse of the scaling operation [Eq. (1)] and retransformed to have units of millimeter per day, applying a square operation. Please note that the same scaling factors calculated for the training set were used for scaling both the generated training and validation sets.
- The climatology, standard deviation, and 95th and 99th percentiles for both the generated training and validation datasets in millimeter per day were calculated. The root-mean-square errors of these statistics were computed, with respect to the corresponding quantities for the ERA5 datasets.
- The power spectra of the generated training and validation datasets in millimeter per day were computed, from which the log-spectral distances with respect to the power spectra of the respective true datasets were derived.

The evolution of all the metrics indicates that the generator converges during training. A rapid improvement is observed in the first 50 epochs, followed by a slow, steady improvement, leading to a stable situation between epochs 300 and 400. Importantly, the trend observed in metrics calculated on the validation set closely follows that of the metrics computed on the training set, which allow us to exclude the occurrence of overfitting during the training process. After conducting several sensitivity analysis, considering the absence of evident overfitting, we

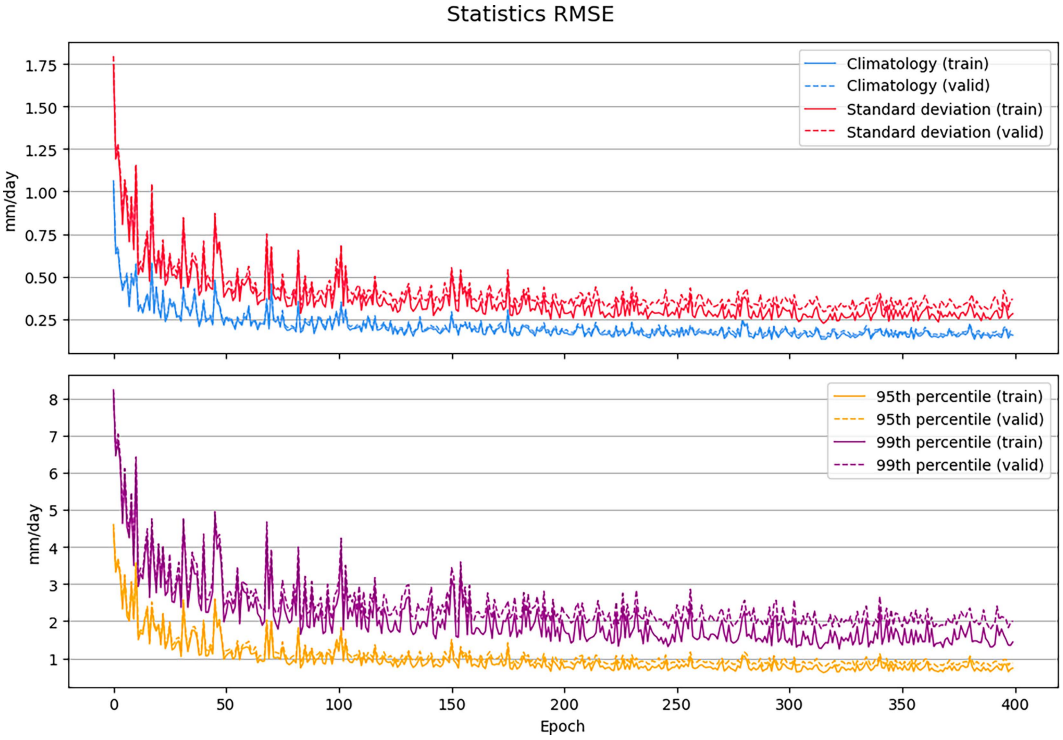

FIG. 5. Evolution of statistical metrics over the GAN training: RMSEs between (top) the climatology and standard deviation, and (bottom) 95th and 99th percentiles of the generated dataset with respect to the corresponding true dataset. The solid lines refer to the training dataset (1940–98), while the dashed lines correspond to the validation dataset (1999–2010).

empirically established that 400 training epochs is the threshold beyond which RainScaleGAN's skill does not significantly improve.

b. *Model selection*

Optimizing RainScaleGAN's hyperparameters is a challenging task, due to the nonmonotonic behavior of the above-defined metrics during the training process (cf. Figs. 4 and 5). This difficulty can be attributed to the intrinsically stochastic nature of the model, stemming from the inclusion of a white noise source in the generator's input. This characteristic does not compromise the quality of the GAN with respect to the downscaling task it is designed for, as a downscaling model of the type developed in this study aims at generating a (set of) possible realization(s) of precipitation at the small scale, having statistical properties that mirror those measured in the corresponding area (Rebora et al. 2006). However, this small-scale stochasticity results in the variability of the generated field, which, while being desirable in studies employing ensembles, may negatively impact grid-point-wise calculated statistics, such as those considered here. Indeed, metrics that compare a forecast and an observation (or, as in this case, a surrogate like a reanalysis) individually in each location suffer from the so-called *double penalty effect* (Rossa et al. 2008): small errors in the placement of the forecasted precipitation result in both the penalty associated with incorrectly locating precipitation (miss) and predicting it in the wrong place (false alarm).

Instead of introducing additional metrics, we opted for a simple selection criterion. We evaluated the model in different configurations, varying one hyperparameter at a time, and selected the configuration that performed best based on the number of epochs—after RainScaleGAN stabilized (i.e., after 300 epochs, as specified in section 3a)—in which it achieved the lowest root-mean-square error across a set of metrics computed for the validation set, relative to the corresponding ERA5 subset. For this task, we considered the four statistical metrics—climatology, standard deviation, and 95th and 99th percentiles—since they effectively evaluate the accurate reproduction of climate, which is our primary interest, and are less prone to excessive variability during the training process. For each of the final 100 epochs in each training run within a given group of model configurations, we identified the configuration that achieved the lowest root-mean-square error for each metric compared to all others in the group. We separately tested the size of the convolutional filters for the generator and the discriminator, varying it between 2×2, 3×3, 4×4, and 5×5. The results of these experiments, depicted in Fig. 6, led us to choose the configuration that achieved the best scores for the most metrics considered, namely, the one with 4×4 convolutional filters for both the generator and the discriminator.

c. *Model evaluation: Analysis of a single realization*

The evaluation of RainScaleGAN was conducted using the test set previously held out for this purpose (years 2011–22). The preprocessing of this dataset followed the same

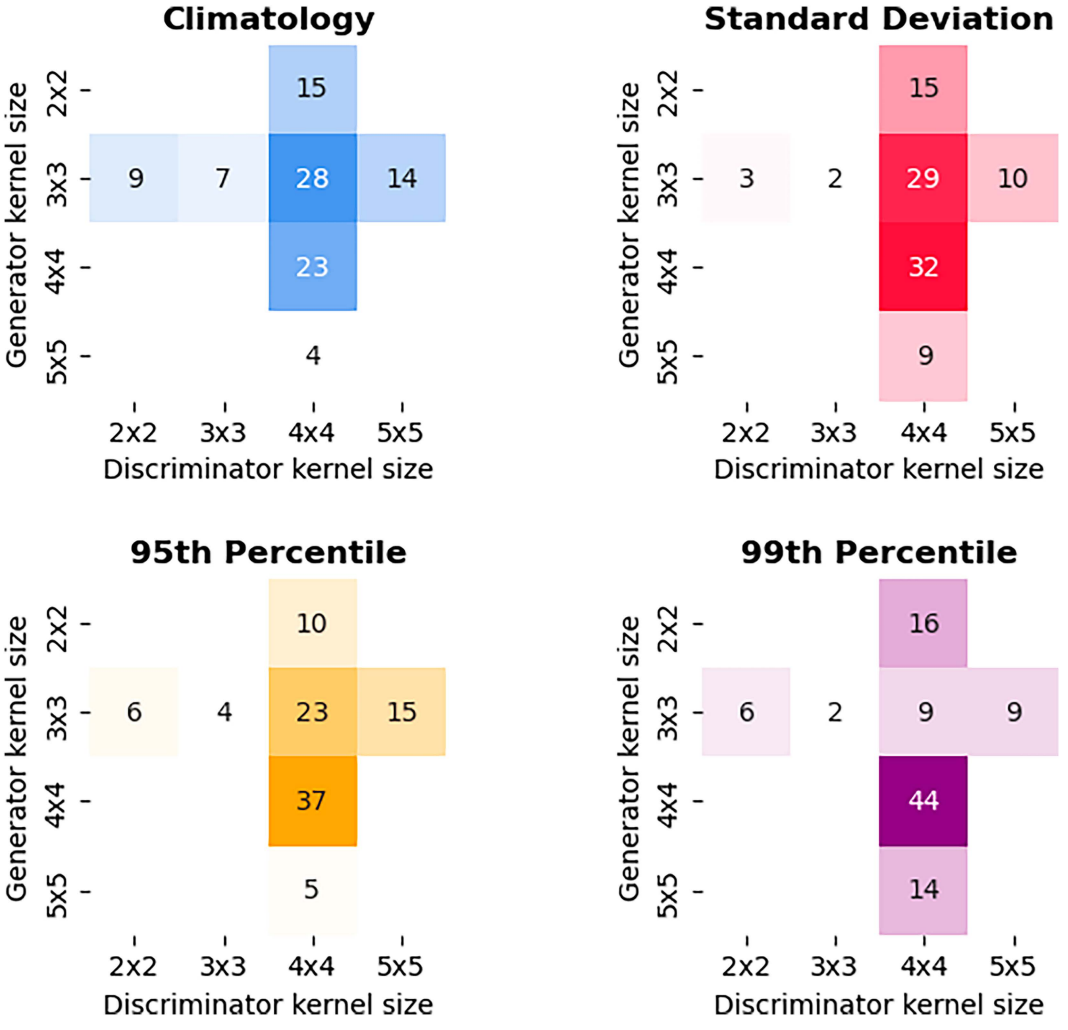

FIG. 6. Selection of convolutional filter sizes. The plot shows, for each configuration and each selection metric considered, the number of training epochs in the final phase of training, after the GAN has stabilized, in which each configuration achieves the lowest RMSE for that metric compared to the other configurations.

procedure applied to the training and validation datasets, as outlined in section 2a. Importantly, the same scaling factors calculated for the training set were used for the scaling of the test set. Since RainScaleGAN is trained to handle rainfall scaled with these factors, the network parameters are adjusted during training to reproduce precipitation magnitudes at each grid point that depend on this scaling. The use of different scaling factors (calculated, for example, on the test set itself) would compromise the generator's ability to reproduce the correct amount of precipitation at each grid point, along with its capacity to extrapolate to previously unseen data. After these operations, the optimal generator identified during the validation phase was used to downscale the entire upscaled test set, which was then denormalized and transformed back to units of millimeter per day. In this way, a dataset was created that should mimic the test set extracted from the original ERA5 data.

Figure 7 shows the comparison between the predictions generated by RainScaleGAN and those of RainFARM, for four randomly selected precipitation events, along with the corresponding ERA5 ground truth. The analysis of the maps highlights that RainScaleGAN produces precipitation fields with more realistic details. While both methods are effective in capturing the large-scale structure of the precipitation field, RainFARM seems constrained to reproduce fine-scale details with the same texture across all parts of the domain. Furthermore, RainFARM's downscaling procedure, while conservative, introduces local maxima at locations distinct from those where the actual maxima are present in the original data. In contrast, RainScaleGAN, despite producing discrepancies compared to the ground-truth field, appears to generate a field that is more visually consistent with the true field. Moreover, it successfully captures the position and magnitude of precipitation maxima, which is particularly desirable in the context of studies on extreme events.

The analysis of the statistical metrics for the test set confirms the superiority of RainScaleGAN. Figure 8 displays the maps

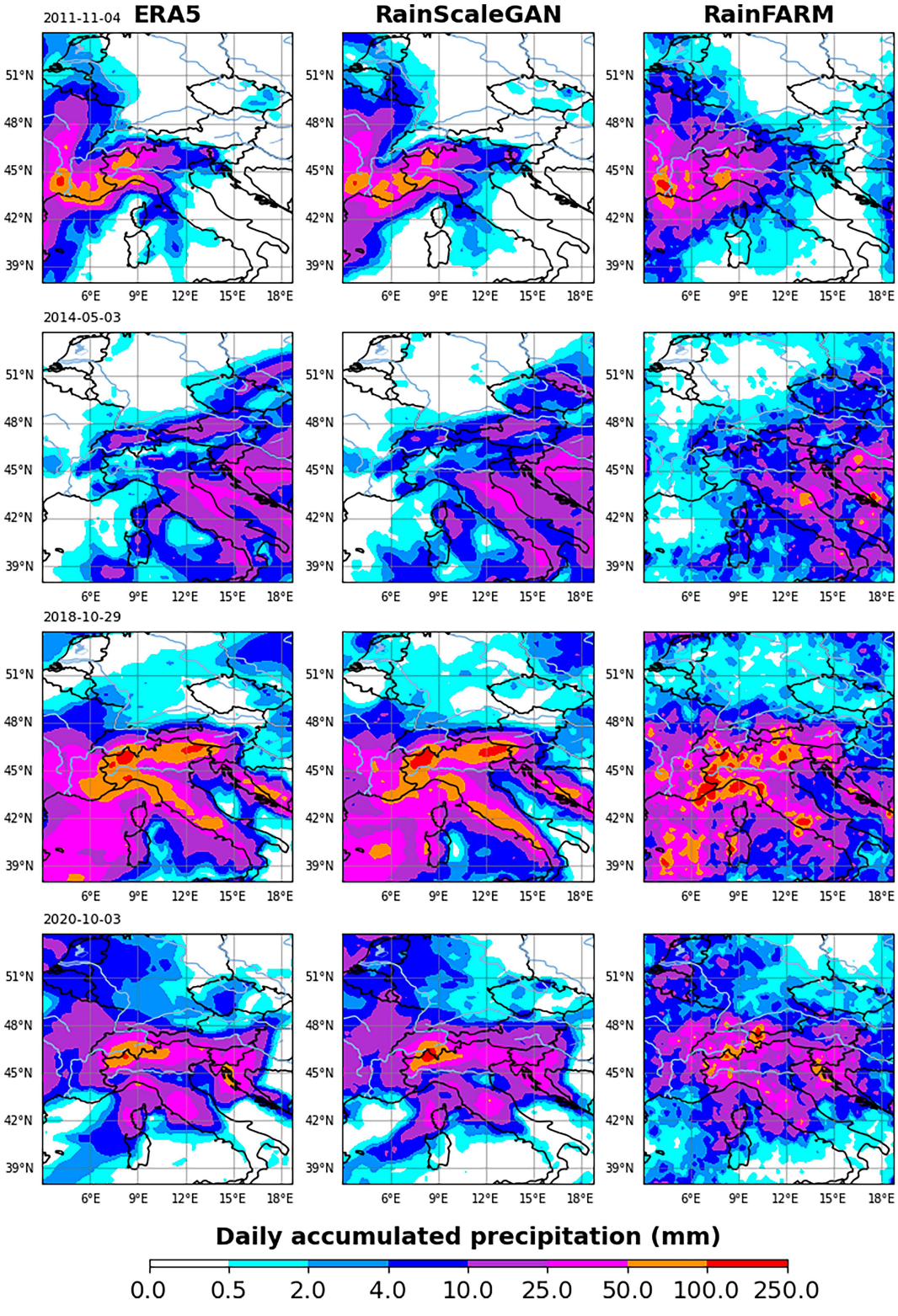

FIG. 7. Comparison between (middle) the predictions generated by RainScaleGAN and (right) the predictions of RainFARM, for four randomly selected precipitation events. (left) The (ERA5) ground-truth data.

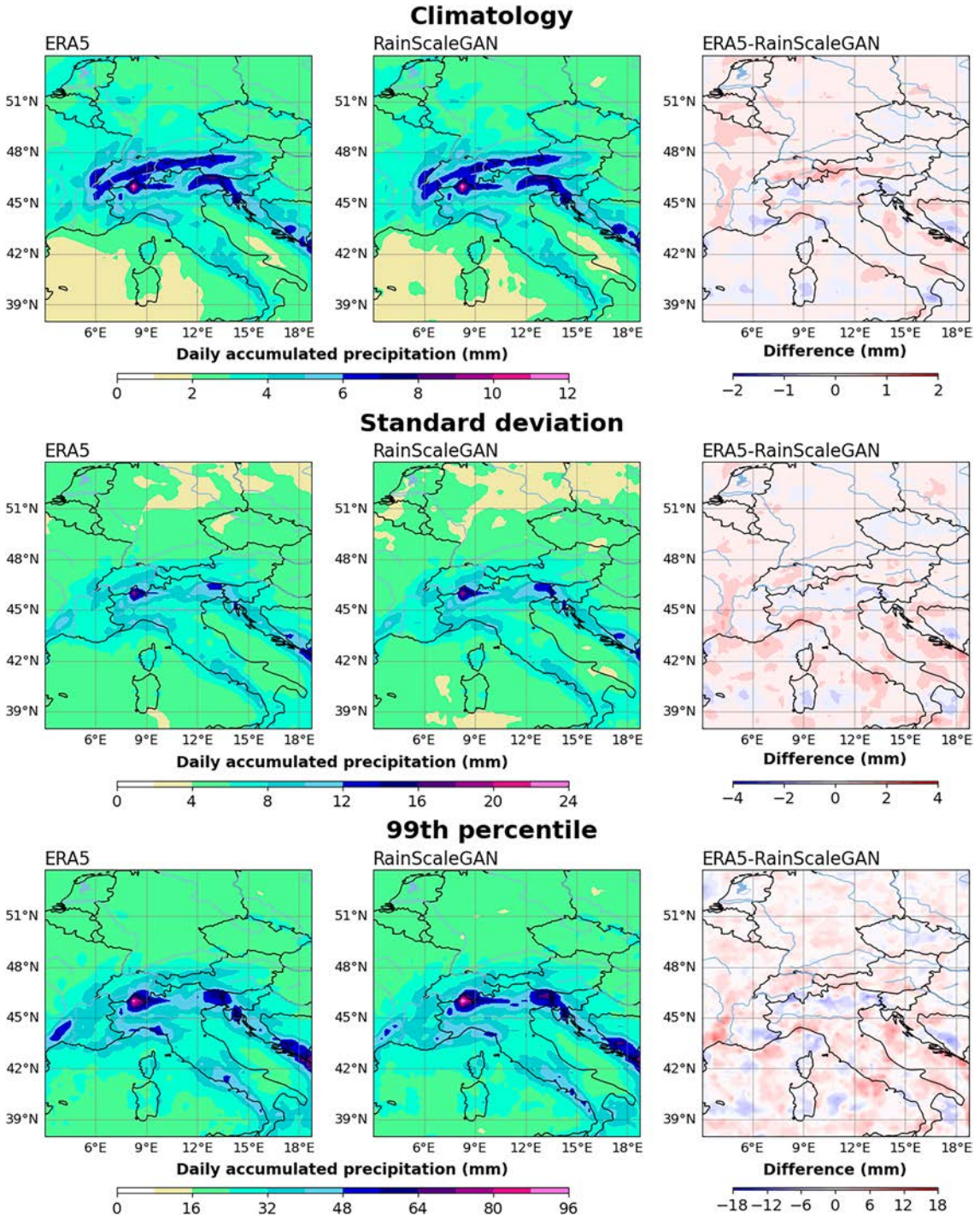

FIG. 8. Maps of statistical metrics for (left) the ERA5 ground-truth test set (2011–22), (middle) those for the test set as reconstructed by RainScaleGAN, and (right) the deviation between the corresponding metrics of the two groups.

of the climatology, standard deviation, and 99th percentile for the test set downscaled by RainScaleGAN. It also includes the metrics for the ERA5 ground-truth dataset, as well as the deviations of these metrics with respect to those of the GAN downscaled dataset. Figure 9 shows similar maps, but considering the test set reconstructed by RainFARM. The observed edge artifacts result from the periodic boundary conditions assumed in its implementation. In operational settings, this issue is typically mitigated by applying the downscaling procedure to a slightly larger domain than the region of interest.

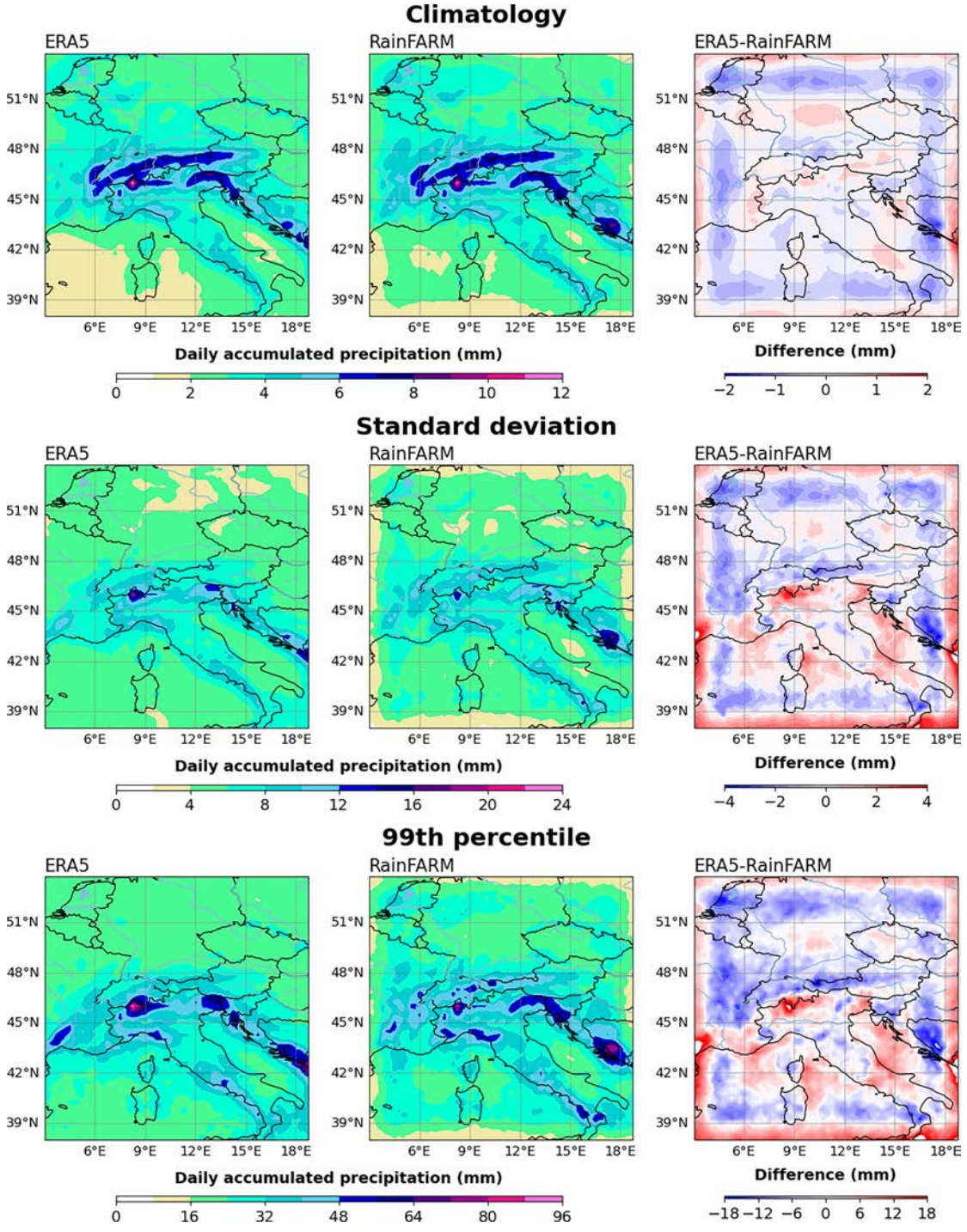

FIG. 9. As in Fig. 8, but for the test set (2011–22) as reconstructed by RainFARM.

Both methods reproduce climatology with sufficient accuracy, correctly capturing its spatial patterns and magnitude. However, RainFARM benefits from incorporating climatological information for the fine-scale precipitation field during calibration, enabling it to refine and correct its predictions accordingly. Although the hold-out test dataset is temporally distinct from the training dataset (used to define the climatology for RainFARM), and thus exhibits different statistical properties, the inclusion of climatological information related to fine-scale precipitation over the extended period corresponding to the training set enhances the RainFARM downscaling. This approach helps capturing the spatial behavior of precipitation, especially in

TABLE 1. RMSEs for the statistical metrics of the downscaled test set, with respect to those of the ERA5 test set.

	RMSE (mm day^{-1})	
	RainScaleGAN	RainFARM
Climatology	0.178 648	0.368 761
Standard deviation	0.389 718	1.121 955
95th percentile	0.920 116	2.443 838
99th percentile	2.111 995	5.848 271

regions with complex orography like the Alps, where the terrain significantly influences the spatial patterns of precipitation. Consequently, this source of information contributes to the observed good outcome. Conversely, the GAN does not have *explicit* access to this type of information. The accurate reproduction of climatology in areas with complex orography suggests that RainScaleGAN, by seeing during the training process examples of precipitation fields sampled from the same probability distribution it aims to reconstruct, is able to autonomously infer its statistical characteristics, including climatology. This explains the excellent results, even without the explicit constraint provided to RainFARM. We consider this a remarkable achievement. For the other statistics, RainScaleGAN continues to excel in reconstruction accuracy, whereas RainFARM introduces artifacts and distortions in both the placement of local maxima and the prediction of their correct magnitudes. Table 1 reports the root-mean-square errors between the statistics of the datasets generated by RainScaleGAN and RainFARM, with respect to the corresponding ERA5 dataset, further confirming the above observations.

Figure 10 displays the time mean of radial power spectra for the test set examples, generated by both RainScaleGAN and RainFARM, along with the corresponding reference spectrum for the ERA5 ground truth dataset. The figure legend also includes the log-spectral distances between the two generated power spectra and the ERA5 spectrum. Details for the calculation of these quantities are provided in section 3d. It is evident that RainScaleGAN produces a dataset whose mean spectrum faithfully reflects that of the reference dataset, while RainFARM loses definition at small scales (high k), where its spectrum appears as a simple extrapolation of that at larger spatial scales. This observation is not surprising, considering the theoretical framework of RainFARM (see the appendix for details).

RainScaleGAN does not explicitly model the temporal evolution of precipitation fields. Consequently, the temporal structure of the downscaled outputs is expected to reflect that of the low-resolution inputs. In other words, the temporal consistency of

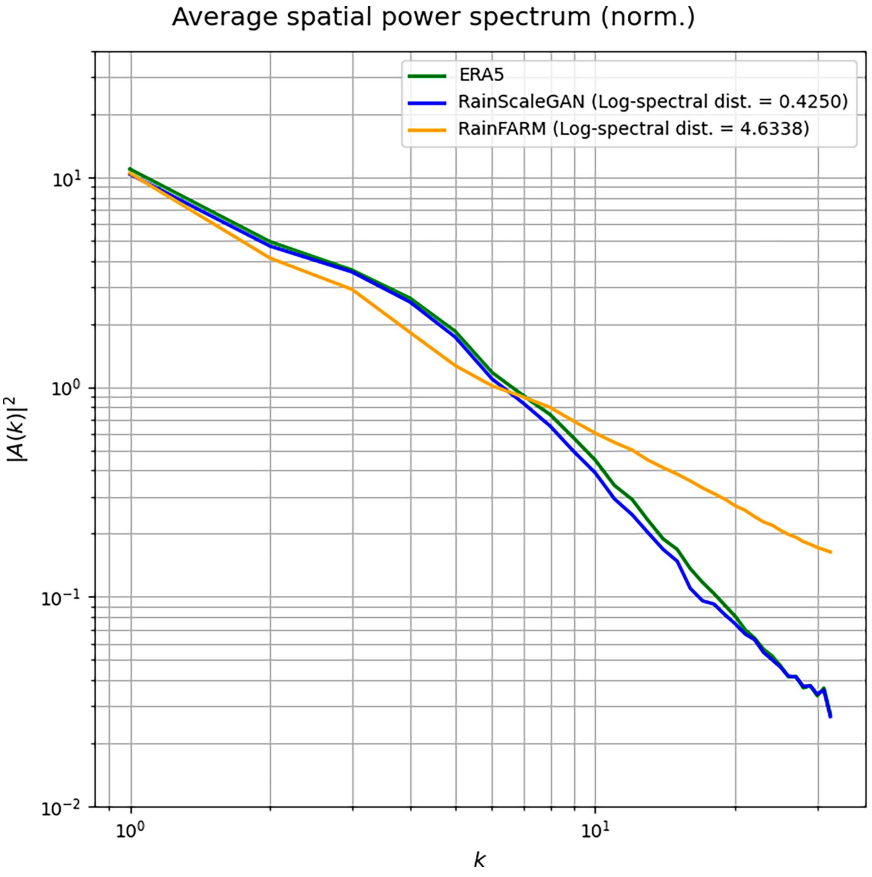

FIG. 10. Time-mean radially averaged power spectra of the downscaled test sets generated by RainScaleGAN and RainFARM, compared with the reference power spectrum of the ERA5 dataset.

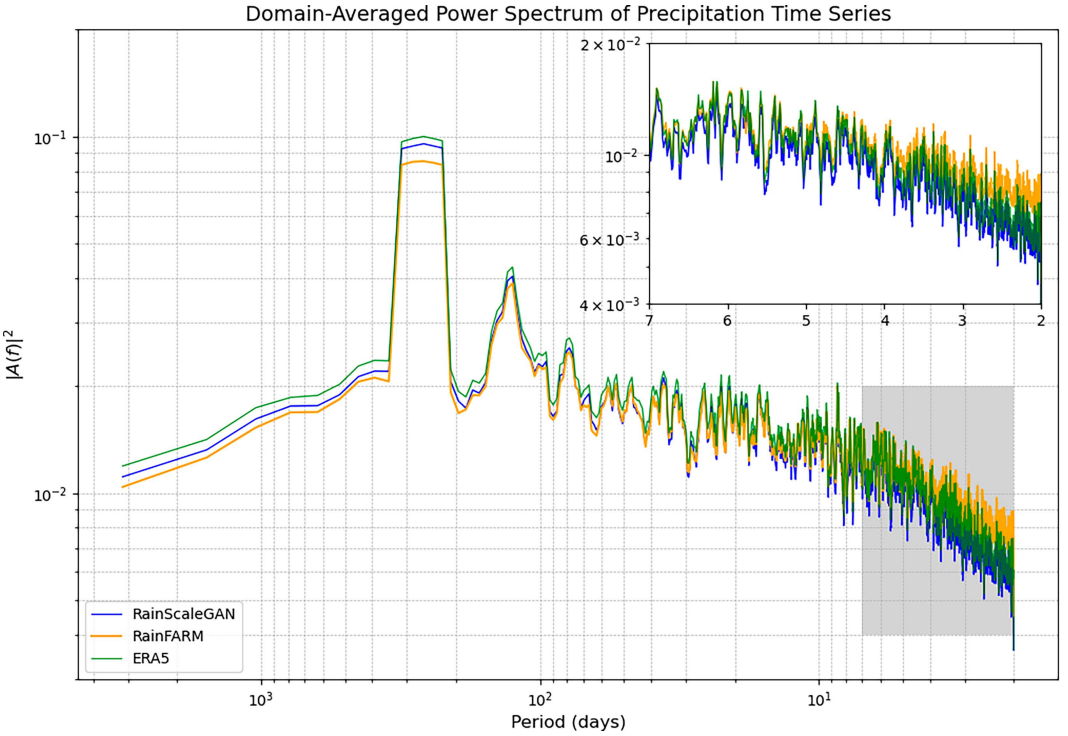

FIG. 11. Domain-averaged power spectrum of the precipitation time series from the downscaled test set generated by RainScaleGAN and RainFARM, compared with the reference power spectrum from the ERA5 dataset.

the generated precipitation relies solely on the model that produces the training data. Nonetheless, it is important to evaluate whether RainScaleGAN distorts the temporal characteristics of precipitation. To investigate this, we analyze the power spectrum of the precipitation time series, averaged over the considered domain. Figure 11 shows the domain-averaged power spectrum of the precipitation time series from the downscaled test set generated by RainScaleGAN and RainFARM, compared to the reference power spectrum from the ERA5 dataset. The power spectrum is computed by applying a normalized Fourier transform to the precipitation time series at each grid point within the test set (2011–22). The resulting spectra are then averaged over the domain. This analysis is performed for ERA5, as well as for a single realization of the precipitation field generated by both RainScaleGAN and RainFARM. To reduce noise, the power spectrum is smoothed using a running mean filter with a window size of 5. A comparison of the three spectra reveals that RainFARM tends to overestimate power at short periods, suggesting an excess of variability at high frequencies. In contrast, RainScaleGAN more closely follows the ERA5 reference spectrum, indicating a more realistic reconstruction of the temporal structure of precipitation variability.

As an additional evaluation tool to assess the downscaling skills of RainScaleGAN, we analyzed the probability distribution of the daily accumulated total precipitation across the entire domain, over the full temporal extent of the test set. We treated all grid points together, considering all time steps, to construct a single probability distribution. Figure 12 shows the complementary cumulative distribution function (CDF) of the ground-truth ERA5 test dataset, together with the complementary CDFs of the test sets reconstructed by RainScaleGAN and RainFARM. For a real-valued random variable X evaluated at x, this quantity is defined as

$$\overline{F}_X(x) = P(X > x) = 1 - F_X(x) = 1 - \int_{-\infty}^{x} f_X(t)dt, \qquad (11)$$

where F_X is the CDF of X, f_X is its probability density function, and $\overline{F}_X(x)$ represents the probability of the variable X exceeding the value x. In our specific case, with X being the daily precipitation values from all grid points and all time steps in the test set, this has a good physical meaning, expressing the probability of a certain precipitation value being exceeded across the entire geographical region. The functions plotted in Fig. 12 demonstrate that RainScaleGAN is able to accurately reconstruct the amount of precipitation over the studied domain, even though it slightly underestimates the rightmost tail of the distribution, generating slightly lower precipitation maxima. On the other hand, RainFARM appears to overestimate the highest values of the precipitation distribution, introducing a significant number of unrealistic extreme values, exceeding the true maxima with value well above 200 mm day^{-1}.

The quantile–quantile (Q–Q) plot in Fig. 13 further confirms these observations. It is constructed by plotting the quantiles of the precipitation probability distribution of the test dataset, as reconstructed by RainScaleGAN and RainFARM, against those of the probability distribution of the ground-truth ERA5 dataset. Each point on the plot represents the precipitation

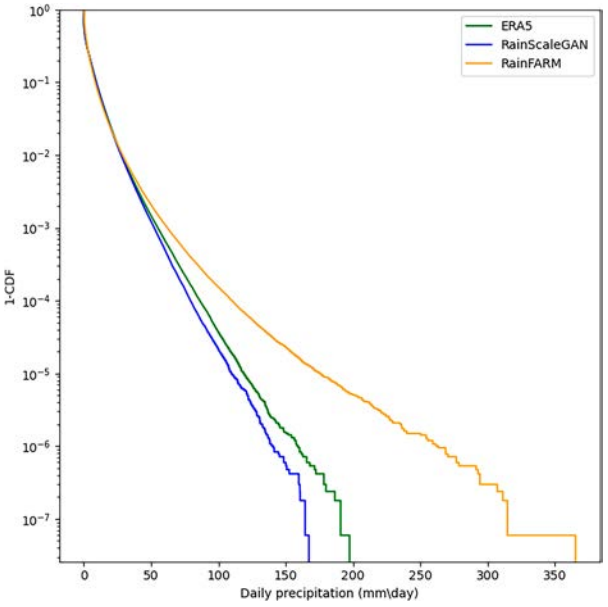

FIG. 12. The complementary CDF of the daily accumulated total precipitation for the ERA5 test set (2011–22), along with the corresponding functions for the test set as reconstructed by RainScaleGAN and RainFARM.

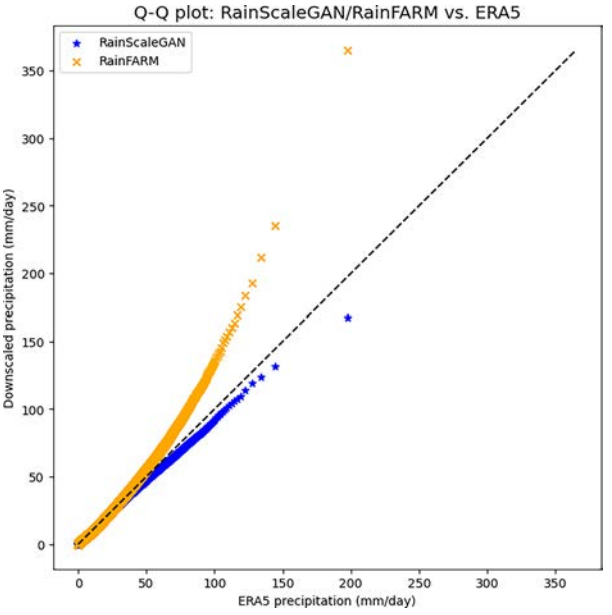

FIG. 13. Q–Q plot of the precipitation probability distribution for the test dataset (2011–22) downscaled by RainScaleGAN and RainFARM against the probability distribution of the ground-truth ERA5 test dataset.

value corresponding to a certain quantile of the probability distribution for the test dataset generated by the two downscaling models against the value of the corresponding quantile for the ground-truth ERA5 test dataset probability distribution. This visualization highlights that while the lowest and central parts of the generated precipitation distribution are satisfactorily captured by both downscaling techniques, RainFARM introduces a positive bias in the uppermost part of the distribution. This is evidenced by the trend of the highest quantiles, which is steeper than the bisector line. Conversely, the trend of the quantiles of the RainScaleGAN downscaled test set closely follows the bisector line, indicating precipitation amounts more consistent with the true amounts across the entire range of values. These observations are consistent with the earlier analysis, as all the statistical metrics of the RainFARM-reconstructed dataset showed positive biases with respect to the corresponding statistics of the ERA5 test set.

d. Model evaluation: Noise impact and reliability

The analysis conducted in section 4c was based on a single realization of the precipitation field at the target scale. While this is important for evaluating the ability of RainScaleGAN to generate a realistic dataset with good statistical properties, there are other aspects of the output generated by the GAN that deserve further investigation. As highlighted in section 3b, the generator includes a noise source—an array of random numbers drawn from a normal distribution with a mean of 0 and a standard deviation of 0.2. Therefore, it is important to explore whether, and to what extent, this noise source influences the results.

To address this, we generated an ensemble of realizations of the precipitation field at the fine-scale, using the test set held out for evaluation (2011–22). The preprocessing of this dataset followed the same steps and precautions (in particular the scaling factors) outlined in section 4c. The same optimal generator identified during the validation phase was used to produce 100 realizations of the upscaled test set. The inclusion of random noise as input to the generator ensures the stochastic nature of the fine-scale details in the output precipitation fields. This approach allows us to investigate the impact of the noise source on RainScaleGAN output.

Figure 14 shows four different realizations of the same set of precipitation events considered in Fig. 7. Comparing these realizations for the same dates provides insight into the variability introduced by the noise input to the generator. An inspection of the maps reveals that while the boundaries and positions of areas associated with precipitation events shift slightly across different realizations, RainScaleGAN appears to be self-consistent in positioning local maxima. This consistency is important for assessing the intensity and location of heavy precipitation events. Importantly, the large-scale spatial structure of the precipitation field is preserved, suggesting that the noise input to the generator does not distort this structure, as prescribed by the coarse-scale field being downscaled.

As in section 4c, to conduct a more quantitative evaluation, we considered the probability distribution for the daily accumulated total precipitation generated by RainScaleGAN by building the complementary CDF for the test set realizations produced by both RainScaleGAN and RainFARM. These results were then compared with the ground-truth ERA5 test set. The corresponding plot is shown in Fig. 15. This plot illustrates the probability of a

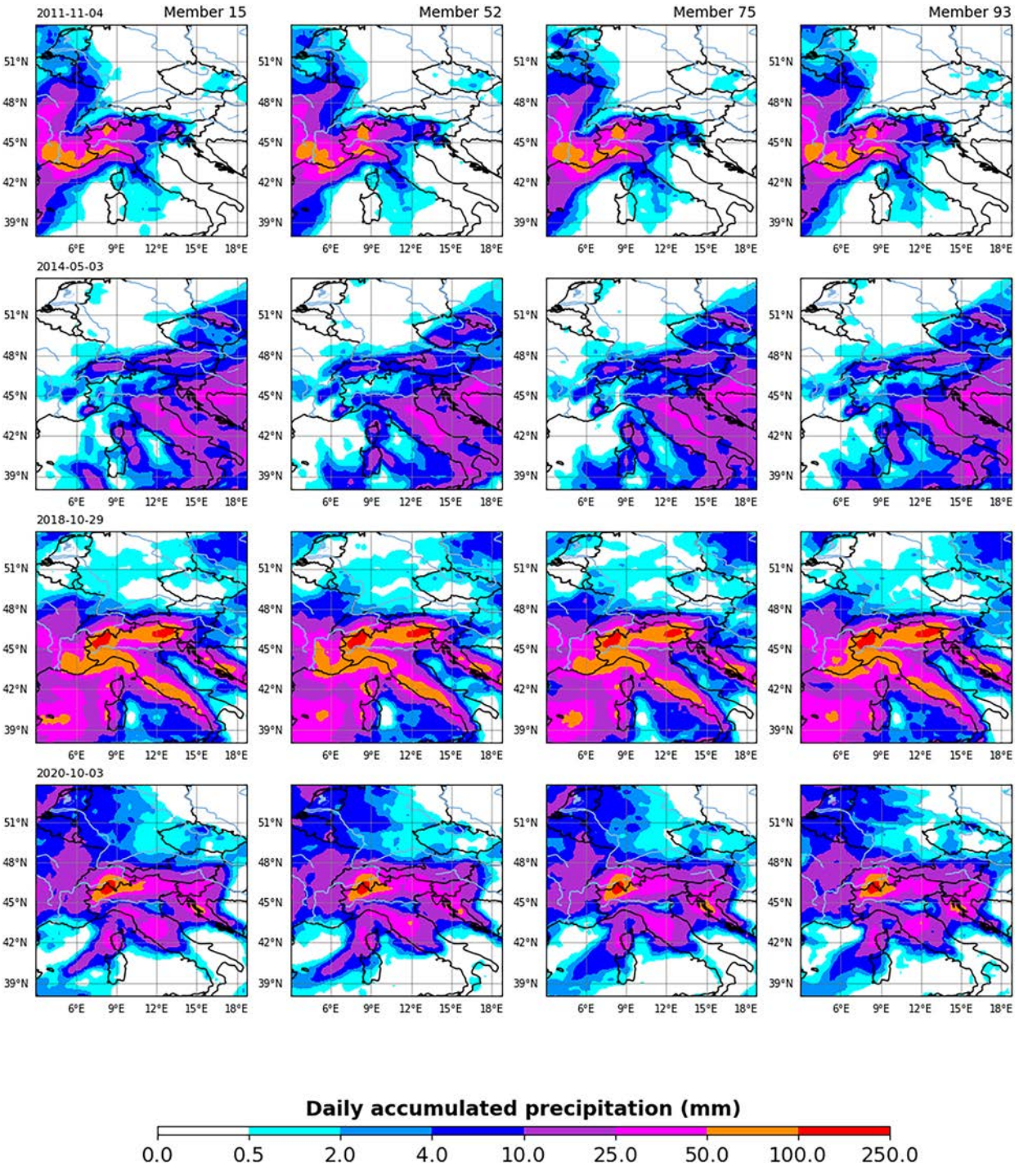

FIG. 14. Four different realizations of the same set of precipitation events shown in Fig. 7.

certain precipitation value being exceeded across the entire considered domain. By inspecting it, we gain insight into the spread of the precipitation maxima produced by both downscaling methods. RainScaleGAN generates a narrower range of precipitation maxima, and, consistent with what was noted in the previous section, it slightly underestimates the rightmost tail of the distribution, yielding lower values. In contrast, RainFARM tends to overestimate the highest values, with some members generating precipitation maxima more than double the ground-truth global maximum for the accumulated precipitation within the investigated domain. While RainScaleGAN appears to be more accurate between the two downscaling

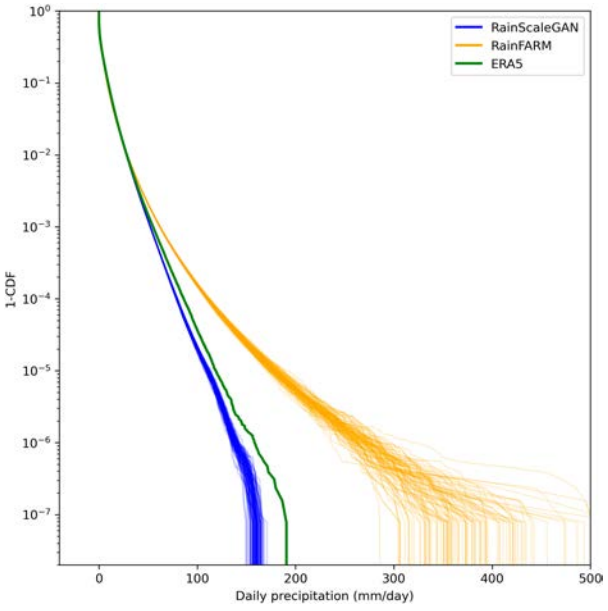

FIG. 15. The complementary CDF of the daily accumulated total precipitation for the ERA5 test set (2011–22), alongside the corresponding functions for 100 realizations of the test set generated by RainScaleGAN and RainFARM.

methods, its restricted range of values may be a disadvantage in studies where a good representation of variability is important. This aspect requires further investigation, involving an assessment of the variability of results as a function of parameters of the noise input to the generator.

Reliability diagrams are a commonly used tool for verifying probabilistic forecasts. In these diagrams, forecasts are grouped into bins according to their predicted probability, shown on the horizontal axis, while the corresponding observed frequency is plotted on the vertical axis. For perfect reliability, the forecast probability should match the observed frequency, resulting in points lying along the diagonal. For a detailed explanation of their construction, interpretation, and meaning, refer to Wilks (2011).

Reliability diagrams are typically used to evaluate probabilistic forecasts for dichotomous events, such as rain versus no rain. Although each realization of the precipitation field from RainScaleGAN and RainFARM can be viewed as a deterministic prediction of a continuous variable, by generating an ensemble of forecasts from the same predictor (the large-scale precipitation field to be downscaled), we can adapt this verification tool to our context. To assess how RainScaleGAN performs across different precipitation intensities—particularly in predicting weak (drizzle) and extreme precipitation events—and to identify any potential biases, we defined three binary events:

1) Total accumulated precipitation < 1 mm day^{-1} (drizzle event).
2) Total accumulated precipitation exceeding the 95th percentile.
3) Total accumulated precipitation exceeding the 99th percentile.

For each time step in the test set, we evaluated the occurrence of these events at each grid point. The thresholds for the 95th and 99th percentiles were determined from the full time series (years 2011–22) at each grid point. These conditions define masks that transform the generated dataset into a binary prediction (yes/no) for the corresponding event. This procedure was applied separately to each ensemble member, after which the ensemble mean was computed at each time step and grid point to derive a single probabilistic prediction of event occurrence. Using the ERA5 ground-truth test set as a reference, we then computed a reliability diagram for each grid point. Finally, we averaged these diagrams spatially across the domain to obtain the domain-averaged reliability diagram. The results of this process are shown in Fig. 16. To complement the reliability diagrams, we also constructed forecast frequency histograms (sharpness diagrams), which illustrate the distribution of forecast probabilities by showing the relative frequency of instances within each probability bin.

The reliability diagrams reveal distinct behaviors for the two models. For RainScaleGAN, reliability varies across precipitation thresholds. For drizzle events, the model exhibits good calibration, with its reliability curve closely following the diagonal, indicating that predicted probabilities align well with observed frequencies. For extreme precipitation, the model becomes increasingly overconfident, overestimating the occurrence of extreme events while maintaining good reliability for low forecast probabilities. This effect is more pronounced at the 99th percentile, where deviations from the diagonal are larger. The sharpness diagram for drizzle shows peaks at the extreme probability bins (near 0 and 1), suggesting that the model most often assigns either very low or very high probabilities. For extreme precipitation, the distribution shifts, with a dominant peak in the first probability class (0–0.1), indicating that the model assigns low probabilities most of the time. However, when it does predict high probabilities, it tends to be overly confident, overestimating extreme events. This suggests that the model is generally cautious in predicting extreme precipitation but overconfident when it does.

The sharpness diagrams for RainFARM exhibit a similar pattern, with peaks at the two extreme probability bins for drizzle events and a single peak in the lowest probability class for extreme events. These patterns indicate that RainFARM, like RainScaleGAN, assigns strong probabilities to its forecasts. The analysis of the reliability curves reveals interesting features. For drizzle events, RainFARM shows a dry bias, with its reliability curve lying above the diagonal. This implies that the actual frequency of drizzle is consistently higher than the predicted probability, indicating that the model systematically underpredicts drizzle events. For the 95th percentile, the reliability curve is slightly steeper than the diagonal, suggesting underconfidence. For the 99th percentile, the curve takes a reverse U shape, indicating that RainFARM is highly overconfident in assigning high probabilities to extreme precipitation events, resulting in overprediction.

The analysis presented in this section does not aim to provide a comprehensive evaluation of the ensemble generated by RainScaleGAN. Instead of a full assessment—which would require the use of ensemble-specific metrics and a quantitative

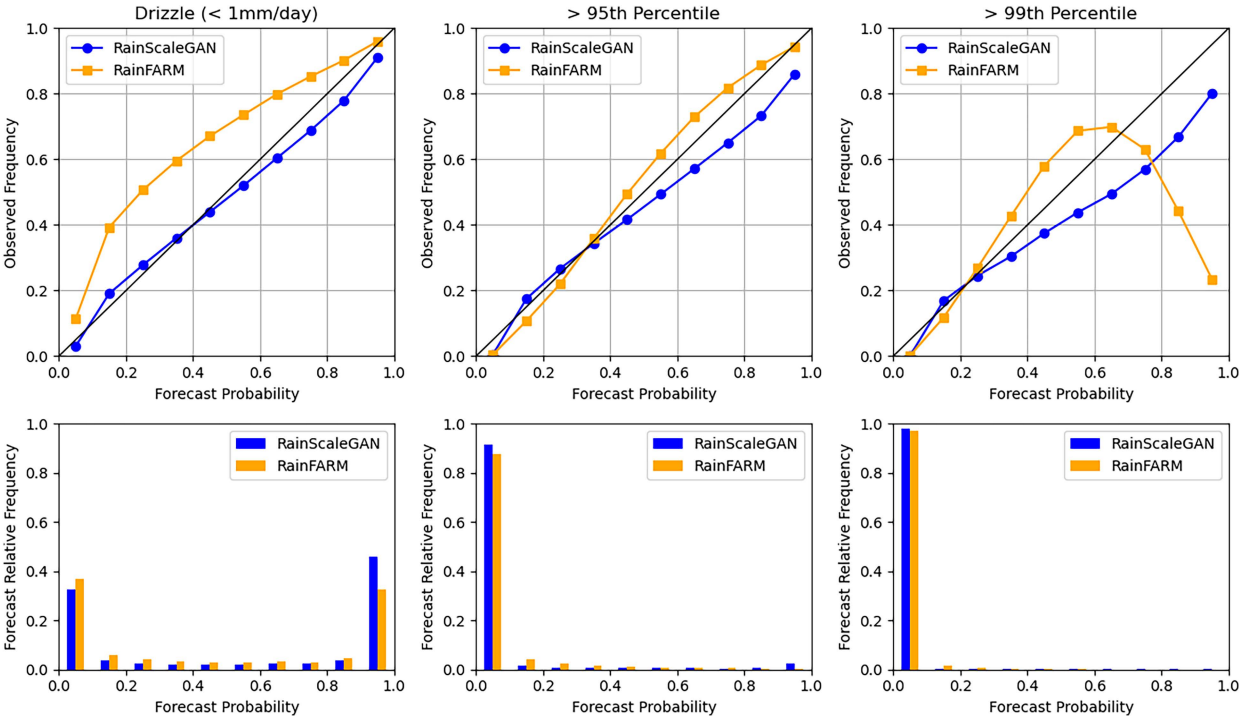

FIG. 16. (top) Domain-averaged reliability diagrams for predicted drizzle (<1 mm day^{-1}) and extreme events (>95th percentile, >99th percentile) for both RainScaleGAN and RainFARM. (bottom) The corresponding forecast frequency histograms (sharpness diagrams), showing the relative frequency of forecast instances in each probability bin.

analysis of the sensitivity of the produced fields to the random noise input—the goal is to illustrate RainScaleGAN potential for generating consistent sets of precipitation field realizations while highlighting the role of noise in shaping the results. In particular, Figs. 14 and 15 emphasize the variability introduced by the generator at small spatial scales. This variability is crucial for precipitation downscaling, as the small-scale features of precipitation events can exhibit significant differences even under the same large-scale conditions.

5. Discussion and conclusions

In this study, we introduced RainScaleGAN, a Wasserstein GAN with a simple architecture specifically tailored for downscaling precipitation in climate studies. We evaluated the performance of our model against RainFARM, a state-of-the-art stochastic downscaling method. Our results demonstrated that the finely tuned GAN can effectively perform downscaling in a perfect-model experiment using daily precipitation data from the ERA5 reanalysis. The generated daily precipitation fields, when considered individually, have a more plausible appearance compared to those produced by the alternative method. Additionally, the reconstructed dataset exhibits climate-related statistical properties that closely reflect those of the ground-truth counterpart.

The model selection part of the downscaling exercise, though challenging due to the peculiarities of the GAN training process, is important for the success of the downscaling task. There is no guarantee that the same hyperparameters will be effective in another geographical region or even for a dataset on a different grid within the same region considered here. Therefore, the proposed validation process must be repeated in these situations to ensure the effective application of the methodology.

The downscaling exercise was conducted between resolutions of $2° \times 2°$ and $0.25° \times 0.25°$, covering scales typical of climate modeling. However, since the proposed methodology does not rely on physical assumptions, there are no a priori limitations on applying the architecture to higher spatial resolutions, targeting storm-scale resolutions relevant for weather prediction.

The spatial resolution of the input field to be downscaled, $2° \times 2°$, is commonly found in climate model projections. This naturally raises the question of RainScaleGAN generalization ability when applied to such projections. Two key factors influence this: 1) the accuracy of the climate model being downscaled and 2) the stationarity of the transfer function implicitly learned by the generator under climate change. The first factor arises because RainScaleGAN is not designed to correct large-scale biases, while the second relates to its assumption, as a statistical downscaling method, that the relationship between large-scale predictors and fine-scale predictands remains stationary. Consequently, a generator trained on coarsened ERA5 data should, in principle, perform similarly to the one in this study when applied to a climate projection. However, verifying this assumption requires further investigation. A sensitivity analysis will be necessary to assess the model performance across different temporal periods and climate scenarios.

From the point of view of statistical downscaling, the currently adopted perfect-model setup, which is standard in the development phase, greatly simplifies the problem. As we pointed out in section 3a, it circumvents the issue related to biases between the predictor and the target dataset by adopting a unique data source. By focusing solely on pure superresolution, this approach can be considered the first step in constructing a proper downscaling system. However, classical statistical downscaling methods consist of predictor–predictand relationships that are calibrated against observations. Therefore, two distinct datasets are involved: one at the low resolution and one at the target resolution. For instance, in an operational context, this could involve the output of a large-scale model simulation and a fine-scale observational dataset. In such cases, the unavoidable presence of biases must be taken into account. As a consequence, the downscaling method is also required to correct these biases, at least to some extent. Even though it is unrealistic to expect major location biases and wrong large-scale patterns to be corrected without improving the large-scale model, correcting the amount of precipitation at the local target scale by nudging it toward observations is a realistic goal. An extension of the methodology described here, which can also address these types of biases, could have significant applications in operational contexts, where it might complement or even replace the need for computationally expensive dynamical downscaling methods based on regional models. In this perspective, we believe our work represents an interesting first step toward this goal.

From the perspective of stochastic downscaling, which aims to generate synthetic time series for meteorological variables, RainScaleGAN shares several advantages with methods in this category and appears to outperform RainFARM, a well-established technique in the field. Like RainFARM, RainScaleGAN relies on a single predictor—the precipitation to be downscaled—without needing additional information at either the large or small scale. In this context, RainScaleGAN's ability to accurately reconstruct precipitation statistics at the target resolution, particularly climatology and higher percentiles, is a notable achievement. The quality of the statistics of the RainScaleGAN-generated precipitation dataset, which surpasses that of the RainFARM-reconstructed dataset, is particularly significant for climatological and hydrological studies. For these applications, the accuracy of these statistics often outweighs the precision of deterministic downscaling of individual precipitation events. Moreover, RainScaleGAN effectively captures local-scale precipitation characteristics influenced by factors such as orography, which impacts its spatial distribution over complex terrains. Orography is a key time-invariant field considered in many downscaling techniques. Some methods incorporate it implicitly (e.g., Mei et al. 2020), others explicitly as a predictor (e.g., Harris et al. 2022), while certain approaches are entirely based on topographic information (e.g., Tesfa et al. 2020; Mital et al. 2022). Unlike these techniques, RainScaleGAN achieves realistic precipitation downscaling without requiring an orographic input. Another strength stemming from the adoption of a single-predictor framework is the potential applicability of the technique to any geographical region, regardless of complex orographic features or land–sea boundaries. Since the model does not rely on explicit geographic or topographic data, its primary limitation in this contest is the quality of the precipitation dataset used for training.

Aiming to devise a stochastic downscaling method justifies the adoption of a conditional GAN (cGAN), a deep learning framework with a higher level of complexity compared to a deterministic neural network. The primary motivation for using a cGAN is its ability to capture the inherent stochasticity of fine-scale precipitation patterns. Unlike deterministic models, which provide a single best-guess estimate, a cGAN can generate multiple plausible downscaled realizations, thus better capturing the variability arising from unresolved subgrid processes. This is particularly relevant for precipitation downscaling, where small-scale features can show significant differences, even under the same large-scale conditions. While generating an ensemble was not the primary goal of this study, we leveraged this capability to construct the reliability diagrams and demonstrate how the noise input to the generator influences the individual downscaled precipitation fields, as well as the probability distribution of daily accumulated total precipitation (cf. Figs. 14 and 15). The intention was not to conduct a full-fledged ensemble-based study, but rather to illustrate the potential of using a stochastic model to generate ensembles at a low computational cost.

The future prospects of this work include testing RainScaleGAN in realistic use cases to verify its effectiveness in bias correction, as discussed above. Another aspect to be explored is the flexibility of RainScaleGAN in its application to different spatial scales, up to the storm scale. In connection with this, quantifying the maximum downscaling factor—the ratio between the spatial resolution of the predictor and the target dataset—is also to be investigated. Additionally, examining the performance of the model across various geographical domains will be essential to assess its robustness in different regions. These investigations will contribute to a comprehensive characterization of the proposed downscaling technique, assessing its potential applicability as a complement or alternative to dynamical downscaling methods. Another aspect worth further analysis is the effect of the noise source on the generator input. The capability to generate an indefinite number of precipitation fields at the target scale, all compatible with the large-scale structure prescribed by the coarse-resolution field, paves the way for studies leveraging ensembles. This holds significant potential for estimating errors in model predictions and offers numerous advantages for climate projections and scenarios, enabling a cost-effective generation of high-resolution precipitation field ensembles. Finally, investigating the extensibility of the proposed method to other meteorological variables, such as temperature, or more intriguingly, wind and humidity, is another prospect that deserves exploration.

Acknowledgments. This paper is based on chapter 3 of the Ph.D. thesis "Exploring Deep Learning-Based Approaches for Precipitation Downscaling," presented by M. Iotti at the University of Bologna in June 2024. M. Iotti sincerely thanks the thesis reviewers, Profs. R. Buizza and C. Pasquero, for their constructive feedback and valuable

insights. He also acknowledges the financial support for his current position from the ICSC – Centro Nazionale di Ricerca in High Performance Computing, Big Data and Quantum Computing, funded by European Union – NextGenerationEU. Hersbach et al. (2023) was downloaded from the Copernicus Climate Change Service (C3S). The results contain modified Copernicus Climate Change Service information. Neither the European Commission nor ECMWF is responsible for any use that may be made of the Copernicus information or data it contains. The authors acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources and support. This work has received funding from the Italian Ministry of Education, University and Research (MIUR) through the JPI Oceans and JPI Climate "Next Generation Climate Science in Europe for Oceans" ROADMAP Project (D. M. 593/2016).

Data availability statement. The code used for training, validation, and testing of RainScaleGAN can be found in the GitHub repository: https://github.com/MclTTI/RainScaleGAN.

APPENDIX

Description of the RainFARM procedure

The RainFARM approach can be summarized in the following steps:

- The spatial power spectrum of the low-resolution precipitation field P to be downscaled is computed (the procedure can be extended to the temporal component of the precipitation field, which we neglect in this study).
- The spectrum is extrapolated to the small unresolved scales, assuming that it approximately follows a power law.
- A Fourier spectrum with random uniform distributed phases g is generated, encompassing wavenumbers corresponding to unresolved scales. The inversion of this spectrum produces a Gaussian field defined on the small scales, which is then normalized to have unit variance.
- A nonlinear transformation of the small-scale Gaussian field is used to generate a synthetic precipitation field $\tilde{p}$. When using an exponential transformation, $\tilde{p}$ is lognormal.
- $\tilde{p}$ is constrained to match the low-resolution field P when aggregated to the resolved scales. This alignment is achieved through the definition of suitable weighting factors, to be applied to $\tilde{p}$.

This procedure is inherently stochastic, as varying the phases of the Fourier spectrum g results in small-scale variations of the outcomes.

Terzago et al. (2018) made an additional refinement to the RainFARM procedure, introducing a method to obtain more realistic fine-scale patterns of precipitation. The goal of this adjustment is to enhance the applicability of RainFARM in climatological and hydrological applications. Additionally, it aims to improve its ability to capture extreme events, particularly in regions characterized by complex orography. The method relies on the availability of a fine-scale precipitation climatology, from which corrective weights for the downscaled field are derived. The precipitation datasets downscaled with this enhanced version of RainFARM exhibit significant improvements in climatology, featuring a greater presence of fine-scale details not obtainable with the standard version, as well as enhancements in the spatial detail, placement, and magnitude of extreme values.

REFERENCES

Annau, N. J., A. J. Cannon, and A. H. Monahan, 2023: Algorithmic hallucinations of near-surface winds: Statistical downscaling with generative adversarial networks to convection-permitting scales. *Artif. Intell. Earth Syst.*, **2**, e230015, https://doi.org/10.1175/AIES-D-23-0015.1.

Arjovsky, M., S. Chintala, and L. Bottou, 2017: Wasserstein GAN. arXiv, 1701.07875v3, https://doi.org/10.48550/arXiv.1701.07875.

Copernicus Climate Change Service (C3S), 2023: ERA5 hourly data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), accessed 31 July 2024, https://doi.org/10.24381/cds.adbb2d47.

Dibike, Y. B., and P. Coulibaly, 2005: Hydrologic impact of climate change in the Saguenay watershed: Comparison of downscaling methods and hydrologic models. *J. Hydrol.*, **307**, 145–163, https://doi.org/10.1016/j.jhydrol.2004.10.012.

D'Onofrio, D., E. Palazzi, J. von Hardenberg, A. Provenzale, and S. Calmanti, 2014: Stochastic rainfall downscaling of climate models. *J. Hydrometeor.*, **15**, 830–843, https://doi.org/10.1175/JHM-D-13-096.1.

Ferraris, L., S. Gabellani, N. Rebora, and A. Provenzale, 2003: A comparison of stochastic models for spatial rainfall downscaling. *Water Resour. Res.*, **39**, 1368, https://doi.org/10.1029/2003WR002504.

Feser, F., B. Rockel, H. von Storch, J. Winterfeldt, and M. Zahn, 2011: Regional climate models add value to global model data: A review and selected examples. *Bull. Amer. Meteor. Soc.*, **92**, 1181–1192, https://doi.org/10.1175/2011BAMS3061.1.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, 2014: Generative adversarial nets. *Advances in Neural Information Processing Systems 27*, NeurIPS, https://papers.nips.cc/paper_files/paper/2014/hash/f033ed80deb0234979a61f95710dbe25-Abstract.html.

——, Y. Bengio, and A. Courville, 2016: *Deep Learning.* The MIT Press, 775 pp.

——, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, 2020: Generative adversarial networks. *Commun. ACM*, **63**, 139–144, https://doi.org/10.1145/3422622.

Gulrajani, I., F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, 2017: Improved training of Wasserstein GANs. arXiv, 1704.00028v3, https://doi.org/10.48550/arXiv.1704.00028.

Harris, L., A. T. T. McRae, M. Chantry, P. D. Dueben, and T. N. Palmer, 2022: A generative deep learning approach to stochastic downscaling of precipitation forecasts. *J. Adv. Model. Earth Syst.*, **14**, e2022MS003120, https://doi.org/10.1029/2022MS003120.

Hersbach, H., and Coauthors, 2023: ERA5 hourly data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), accessed 31 July 2024, https://doi.org/10.24381/cds.adbb2d47.

Ioffe, S, and C. Szegedy, 2015: Batch normalization: Accelerating deep network training by reducing internal covariate shift. *Proc. 32nd Int. Conf. on Machine Learning*, Lille, France, JMLR.org, 448–456, https://dl.acm.org/doi/10.5555/3045118.3045167.

Kingma, D. P., and J. Ba, 2017: Adam: A method for stochastic optimization. arXiv, 1412.6980v9, https://doi.org/10.48550/arXiv.1412.6980.

Kumar, B., R. Chattopadhyay, M. Singh, N. Chaudhari, K. Kodari, and A. Barve, 2021: Deep learning–based downscaling of summer monsoon rainfall data over Indian region. *Theor. Appl. Climatol.*, **143**, 1145–1156, https://doi.org/10.1007/s00704-020-03489-6.

Ledig, C., and Coauthors, 2017: Photo-realistic single image super-resolution using a generative adversarial network. *2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, Institute of Electrical and Electronics Engineers, 105–114, https://doi.org/10.1109/CVPR.2017.19.

Leinonen, J., D. Nerini, and A. Berne, 2021: Stochastic super-resolution for downscaling time-evolving atmospheric fields with a generative adversarial network. *IEEE Trans. Geosci. Remote Sens.*, **59**, 7211–7223, https://doi.org/10.1109/TGRS.2020.3032790.

Maraun, D., and Coauthors, 2010: Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Rev. Geophys.*, **48**, RG3003, https://doi.org/10.1029/2009RG000314.

Mei, Y., V. Maggioni, P. Houser, Y. Xue, and T. Rouf, 2020: A nonparametric statistical technique for spatial downscaling of precipitation over High Mountain Asia. *Water Resour. Res.*, **56**, e2020WR027472, https://doi.org/10.1029/2020WR027472.

Mirza, M., and S. Osindero, 2014: Conditional generative adversarial nets. arXiv, 1411.1784v1, https://doi.org/10.48550/arXiv.1411.1784.

Mital, U., D. Dwivedi, J. B. Brown, and C. I. Steefel, 2022: Downscaled hyper-resolution (400 m) gridded datasets of daily precipitation and temperature (2008–2019) for the East–Taylor subbasin (western United States). *Earth Syst. Sci. Data*, **14**, 4949–4966, https://doi.org/10.5194/essd-14-4949-2022.

Piani, C., J. O. Haerter, and E. Coppola, 2010: Statistical bias correction for daily precipitation in regional climate models over Europe. *Theor. Appl. Climatol.*, **99**, 187–192, https://doi.org/10.1007/s00704-009-0134-9.

Price, I., and S. Rasp, 2022: Increasing the accuracy and resolution of precipitation forecasts using deep generative models. *Proc. 25th Int. Conf. on Artificial Intelligence and Statistics*, Valencia, Spain, PMLR, 10 555–10 571, https://proceedings.mlr.press/v151/price22a/price22a.pdf.

Ravuri, S., and Coauthors, 2021: Skilful precipitation nowcasting using deep generative models of radar. *Nature*, **597**, 672–677, https://doi.org/10.1038/s41586-021-03854-z.

Rebora, N., L. Ferraris, J. von Hardenberg, and A. Provenzale, 2006: RainFARM: Rainfall downscaling by a filtered autoregressive model. *J. Hydrometeor.*, **7**, 724–738, https://doi.org/10.1175/JHM517.1.

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

Rossa, A., P. Nurmi, and E. Ebert, 2008: Overview of methods for the verification of quantitative precipitation forecasts. *Precipitation: Advances in Measurement, Estimation and Prediction*, S. Michaelides, Ed., Springer, 419–452, https://doi.org/10.1007/978-3-540-77655-0_16.

Rummukainen, M., 1997: Methods for statistical downscaling of GCM simulations. SMHI Tech. Rep. RMK 80, 44 pp., https://www.smhi.se/download/18.38e7941719209b36a1fc734/1728370724723/RMK_80.pdf.

——, 2010: State-of-the-art with regional climate models. *Wiley Interdiscip. Rev.: Climate Change*, **1**, 82–96, https://doi.org/10.1002/wcc.8.

Schultz, M. G., C. Betancourt, B. Gong, F. Kleinert, M. Langguth, L. H. Leufen, A. Mozaffari, and S. Stadtler, 2021: Can deep learning beat numerical weather prediction? *Philos. Trans. Roy. Soc.*, **A379**, 20200097, https://doi.org/10.1098/rsta.2020.0097.

Sha, Y., D. J. Gagne II, G. West, and R. Stull, 2020: Deep-learning-based gridded downscaling of surface meteorological variables in complex terrain. Part II: Daily precipitation. *J. Appl. Meteor. Climatol.*, **59**, 2075–2092, https://doi.org/10.1175/JAMC-D-20-0058.1.

Terzago, S., E. Palazzi, and J. von Hardenberg, 2018: Stochastic downscaling of precipitation in complex orography: A simple method to reproduce a realistic fine-scale climatology. *Nat. Hazards Earth Syst. Sci.*, **18**, 2825–2840, https://doi.org/10.5194/nhess-18-2825-2018.

Tesfa, T. K., L. R. Leung, and S. J. Ghan, 2020: Exploring topography-based methods for downscaling subgrid precipitation for use in Earth system models. *J. Geophys. Res. Atmos.*, **125**, e2019JD031456, https://doi.org/10.1029/2019JD031456.

Wang, F., D. Tian, L. Lowe, L. Kalin, and J. Lehrter, 2021: Deep learning for daily precipitation and temperature downscaling. *Water Resour. Res.*, **57**, e2020WR029308, https://doi.org/10.1029/2020WR029308.

Wilby, R. L., and T. M L. Wigley, 1997: Downscaling general circulation model output: A review of methods and limitations. *Prog. Phys. Geogr. Earth Environ.*, **21**, 530–548, https://doi.org/10.1177/030913339702100403.

——, L. E. Hay, and G. H. Leavesley, 1999: A comparison of downscaled and raw GCM output: Implications for climate change scenarios in the San Juan River basin, Colorado. *J. Hydrol.*, **225**, 67–91, https://doi.org/10.1016/S0022-1694(99)00136-5.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences.* 3rd ed. Elsevier, 676 pp.