Does Nonstationarity Affect GAN-Based Downscaling? Insights from High-Resolution WRF Simulations over the Canadian Prairies

HEBATALLAH M. ABDELMOATY,[a,b] SIMON MICHAEL PAPALEXIOU,[c,a] ANTONIOS MAMALAKIS,[d,e] SHIVAM SINGH,[d,f] VINCENZO COIA,[g] MELISSA HAIRABEDIAN,[g] PASCAL SZEFTEL,[g] AND PATRICK GROVER[g]

[a] *Department of Civil Engineering, Schulich School of Engineering, University of Calgary, Calgary, Alberta, Canada*
[b] *Irrigation and Hydraulics Department, Faculty of Engineering, Cairo University, Giza, Egypt*
[c] *Institute for Global Water Security, Hamburg University of Technology, Hamburg, Germany*
[d] *Department of Environmental Sciences, University of Virginia, Charlottesville, Virginia*
[e] *School of Data Science, University of Virginia, Charlottesville, Virginia*
[f] *Environmental Institute, University of Virginia, Charlottesville, Virginia*
[g] *BGC Engineering Inc., Vancouver, British Columbia, Canada*

(Manuscript received 30 September 2025, in final form 9 February 2026, accepted 23 March 2026)

ABSTRACT: Although downscaling models are typically trained on historical data and then applied to future projections, this practice raises concerns about their ability to reproduce nonstationary climate signals. While such limitations are expected for traditional statistical approaches, deep learning models have the potential to learn a more generalizable downscaling relationship, separating the transformation task from the nonstationary signal. In this study, we examine the sensitivity of deep learning–based downscaling to nonstationarity using Wasserstein generative adversarial networks (WGANs) to downscale hourly precipitation from 50 to 4 km over the Canadian Prairies. We trained two models on convection-permitting Weather Research and Forecasting (WRF) simulations: one on a historical period (2000–15) and another on a future period (2086–2100). Four setups were considered: hh (trained/tested on historical), ff (trained/tested on future), fh (trained on future/tested on historical), and hf (trained on historical/tested on future). Within-climate models (hh and ff) reproduced spatial heterogeneity more effectively, with better spatial correlations and realistic variability, while all four reproduced temporal autocorrelation. Extremes remain a challenge, yet with negative biases, in terms of hourly precipitation upper percentiles (75th, 95th, 99th, and 99.9th) and daily maxima. The hh and ff cases performed best; the fh case benefited from exposure to stronger extremes during training, while hf performed poorest due to weak generalization. Importantly, however, cross-climate model biases (hf and fh) were not substantially larger than within-climate ones. These findings highlight the challenge of training on stationary data for nonstationary climates but also demonstrate the potential of WGANs to mitigate such limitations.

KEYWORDS: Downscaling; Nonstationarity; Regional models; Deep learning

1. Introduction

Climate models known as general circulation models (GCMs) are designed to simulate large-scale interactions within Earth's climate system, including the distribution of water and energy around the globe. The most recent generation, developed under the CMIP6 initiative, has shown notable improvements over its predecessor, CMIP5, particularly in simulating overall precipitation amounts (Chen et al. 2020; Luo et al. 2022; Song et al. 2021). Despite this progress, GCMs still face challenges in capturing the finer details of precipitation patterns, especially localized, short-duration events like convective storms, due to their relatively coarse spatial resolution (Chen et al. 2021; Demott et al. 2007; Yang and He 2021). This limitation has important implications for infrastructure planning and management, as underestimating the intensity and frequency of such events can lead to inadequate design standards and increased vulnerability to flooding and stormwater impacts. To mitigate these resolution limitations, climate models use parameterization techniques to approximate the effects of small-scale phenomena, such as convection, which occur at the subgrid scale (Gregory 1997; Jakob and Klein 2000). However, because these approximations depend on simplified physics, empirical data, and observational tuning, they often introduce uncertainties, particularly in regions dominated by convective activity. Consequently, the limited spatial resolution of GCM outputs restricts their direct use in local-scale impact assessments and decision-making.

To bridge the gap between the coarse resolution of GCMs and the fine-scale information needed for local applications, downscaling methods are essential. These approaches generally fall into two categories: physical downscaling, which employs regional models to simulate finer-scale processes; and statistical downscaling, which uses data-driven methods to relate large-scale predictors to local climate responses. One widely used physical approach is dynamic downscaling, which utilizes GCM outputs to develop regional climate models (RCMs) or convection-permitting models (CPMs) that simulate physical

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/AIES-D-25-0093.s1.

Corresponding author: Hebatallah M. Abdelmoaty, heba.abdelmoaty@ucalgary.ca

DOI: 10.1175/AIES-D-25-0093.1 e250093

© 2026 American Meteorological Society. This published article is licensed under the terms of the default AMS reuse license. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

Unauthenticated | Downloaded 10/09/26 09:56 AM UTC

processes at spatial scales finer than those of GCMs. RCMs typically operate at resolutions with a grid dimension between 10 and 50 km, while CPMs resolve atmospheric processes at even finer scales, down to 4–10 km, making them capable of explicitly representing mesoscale phenomena such as convective systems, yet at a high computational cost (Chan et al. 2013; Dickinson 1995). Hence, outputs from CPMs are limited in the space and time domains. The Weather Research and Forecasting (WRF) Model is a state-of-the-art CPM designed for both atmospheric research and operational forecasting applications. Developed by the National Center for Atmospheric Research (NCAR) and a community of collaborators, WRF offers a flexible, modular framework that supports a wide range of applications, from idealized simulations to high-resolution regional climate modeling (Skamarock et al. 2005). WRF outputs are primarily available at 4 km × 4 km resolution across various regions in North America.

Recent innovations in statistical downscaling have been motivated by the integration of deep learning approaches that can capture the complex spatial and temporal dynamics inherent in climate data, particularly the fine spatiotemporal precipitation fields. These methods aim to mitigate the resolution of coarse GCM outputs by learning sophisticated mappings to fine-scale atmospheric patterns. Deep learning architectures comprise multiple interconnected layers that excel at identifying hierarchical structures in high-dimensional datasets (Liu et al. 2017). Among these layers, convolutional neural networks (CNNs) have proven particularly effective for spatial analysis, as they can detect localized features, textures, and spatial dependencies within gridded datasets (Albawi et al. 2017; Bouvrie 2006; O'Shea and Nash 2015). When time-dependent processes like evolving precipitation patterns are considered, the use of convolutional long short-term memory (ConvLSTM) layers becomes essential. These hybrid structures leverage the spatial processing strength of CNNs with the sequential learning capability of recurrent networks, making them ideal for modeling weather dynamics.

Generative adversarial networks (GANs) are designed to synthesize data samples by training a generator and a discriminator within a competitive framework (i.e., adversarial). They have been proved to perform well in downscaling various hydroclimatic variables (Goodfellow et al. 2016; Iotti et al. 2025; Liu et al. 2025). Conditional GANs include a generation process which is restricted to specific inputs, such as low-resolution climate fields, making them particularly suitable for downscaling tasks (Mirza and Osindero 2014). Despite their promise, GANs frequently face obstacles like training instability and limited diversity in outputs (mode collapse; Saxena and Cao 2021; Zou et al. 2023). These challenges are alleviated by the Wasserstein GAN (WGAN) model, which uses a loss function based on the Wasserstein distance to smooth gradient flows and increase the convergence robustness during training (Leinonen et al. 2021). In the WGAN framework, the generator seeks to reconstruct high-resolution fields from coarse inputs, while the critic (in place of a traditional discriminator) evaluates the realism of the outputs by comparing them with actual high-resolution observations. During training, the two networks compete, driving each other toward convergence. This adversarial learning structure is especially beneficial for downscaling precipitation data, which are characterized by sharp gradients, localized extremes, and non-Gaussian variability, features that simpler interpolation or regression methods are unable to replicate accurately (Saxena and Cao 2021).

WGANs have shown strong potential in generating high-fidelity, spatially coherent precipitation fields (Annau et al. 2023; Chaudhuri and Robertson 2020). For example, a WGAN model was used to downscale hourly precipitation from the Integrated Forecasting System (IFS) at a 10-km resolution to the finer 1-km resolution of the NUMROD radar network across mainland United Kingdom during 2016–20. The model was evaluated under various conditions, including extreme rainfall events, and was found to be consistently reliable (Harris et al. 2022). In another study, a WGAN was applied to downscale tropical cyclone precipitation across five major ocean basins (North Atlantic, west Pacific, east Pacific, South Pacific, and south Indian) from 1979 to the present. Using paired samples derived from an observational dataset (i.e., low-resolution inputs aggregated to 100 km and target fields at 10 km), the model successfully captured detailed spatial features, with mean biases within 5% of the observations, as well as showing an excellent skill in reproducing extreme storms (Vosper et al. 2023).

More recently, a conditional WGAN was also employed to refine ERA5 reanalysis data (~25 km; Hersbach et al. 2020) to 1-km resolution using the Multi-Radar Multi-Sensor (MRMS) dataset over the northeastern United States, focusing on the wettest months between 2018 and 2023 in a region highly susceptible to both extreme rainfall and drought. The model showed strong performance across multiple evaluation metrics, effectively capturing hourly precipitation patterns and extremes (Lee and Park 2025). Similarly, Reddy and Matear (2025) demonstrated the potential of a WGAN model to downscale hourly ERA5 precipitation to the 4.4-km resolution over Tasmania, Australia, while explicitly examining sources of uncertainty such as input variability, noise seed, and model generation.

Although downscaling models are typically trained on historical data, given their wide availability, and then applied to future climate projections, this approach inevitably raises concerns about their ability to reproduce nonstationary signals, particularly for extremes. It is well understood that conventional statistical methods perform poorly under such conditions, as they implicitly assume stationarity (e.g., Lanzante et al. 2018; Tefera et al. 2024). However, deep learning–based models can, in principle, learn the downscaling relationship independently of the underlying nonstationary statistics of the field, allowing them to better capture the magnitude of future extremes.

Despite the advances of GAN-based downscaling methods, the question of whether these methods can model nonstationarity in climate fields remains unanswered. In this study, we aim to address this question by training a WGAN model to downscale coarse-resolution fields (50 km) to a finer 4-km resolution using convection-permitting WRF simulations over the Canadian Prairies as a training dataset. The dataset includes 15 years of historical simulations and 15 years of future simulations. Two separate WGAN models were trained: one on the historical period and another on the future period. Each was evaluated on its ability to reproduce various properties of precipitation fields

Unauthenticated | Downloaded 10/09/26 09:56 AM UTC



FIG. 1. The study region within the Canadian Prairies, mainly covering the provinces of Alberta and Saskatchewan.

across both time domains, including wet/dry occurrence probabilities, advanced general statistics (i.e., L-moments), spatial and temporal dependence, distributional properties, and extremes. This experimental design allowed us to investigate whether the downscaling task is sensitive to nonstationarity.

2. Data

a. Study area

In Canada, severe precipitation events commonly occur as intense thunderstorms lasting a few hours during the warmer seasons of spring and summer (Hanesiak et al. 2022; Milrad et al. 2011). These storms can lead to sudden, localized flooding, causing significant disruptions to agricultural productivity, overwhelming drainage systems, and damaging critical infrastructure, especially in smaller watersheds. Moreover, the changing climate is increasing the frequency and intensity of these extreme weather events across Canada, making them a growing concern for sustainable food production, water management, urban planning, and disaster risk reduction (Brooks 2013). Such events highlight the need for improved understanding of the future patterns of subdaily precipitation at the scale of small watersheds. Accordingly, we focus on a region within the Canadian Prairies to investigate the nonstationarity effects on the downscaling task implemented by GAN-based models. The Prairie region, characterized by its semiarid climate and dependence on agriculture, is especially susceptible to the impacts of extreme precipitation. The geographic extent of the study ranges from 121° to 100° west and from 48° to 61° north (Fig. 1).

b. Simulations

For the reference dataset, we use the outputs of the convective-permitting model, WRF, at a spatial resolution of 4 km and an hourly time step over the Canadian Prairies (Li et al. 2019). The WRF simulations were configured over a domain encompassing mountainous and prairie regions, with physics schemes optimized for cold-season processes. Two scenarios are considered: 1) a historical simulation for the period 2001–15 and 2) a pseudoglobal warming (PGW) simulation representing future conditions from 2086 to 2100, in which climate change signals from CMIP5 models (scenario RCP8.5) were added as boundary and initial conditions. The historical WRF output was evaluated against station-based and gridded observational datasets, demonstrating good agreement in capturing the spatial and seasonal patterns of precipitation, though with a consistent wet bias over mountainous areas. These WRF simulations serve as a physically realistic benchmark for downscaling and model evaluation under current and future climate conditions.

The WRF precipitation fields used in this study consist of 214 032 grid cells (588 longitudes × 364 latitudes), covering the study domain at a spatial resolution of approximately 4 km in a single grid dimension. To emulate a coarse resolution of a high-resolution GCM and typical RCM outputs, these fields are spatially averaged and resampled to a coarser grid of 1092 cells (42 longitudes × 26 latitudes), corresponding to a roughly 50-km resolution. The deep learning model is trained to learn the mapping from these low-resolution (50 km) inputs to their high-resolution (4 km) counterparts using the WRF data as the training dataset. The input data comprise hourly precipitation fields structured as 3D tensors (longitude × latitude × 24 hourly time steps); therefore, the model is trained to downscale 1 day at a time. The WGAN model is designed to process these spatiotemporal inputs and reconstruct high-resolution precipitation fields while preserving temporal dynamics, effectively performing superresolution downscaling with the same temporal resolution.

3. Methods

a. Model structure

Recent studies have already demonstrated the advantages of GAN-based models over traditional statistical and alternative deep learning approaches for precipitation downscaling, particularly in their ability to reproduce realistic spatial structure and variability (e.g., Abdelmoaty et al. 2025a; Leinonen et al. 2021; Papalexiou and Mamalakis 2025). Building on these advances, the present study does not aim to benchmark WGANs against baseline models but instead focuses on evaluating their sensitivity and robustness when applied within and across distinct climate periods. Figure 2 illustrates the architecture of the WGAN model used in this study, including the structure of the generator and critic networks. A key feature of this design is the use of residual blocks, each comprising two convolutional layers followed by activation functions. These blocks apply transformations to the input and then add the original input back to the output, allowing the network to deepen while avoiding the issue of gradient shattering (Balduzzi et al. 2018). When the weights in a residual block are close to zero, the block effectively acts as a pass-through, leaving the input unchanged. As training progresses, the model selectively activates more blocks to capture deeper features. Although adding more residual blocks increases computational demands, it does

Unauthenticated | Downloaded 10/09/26 09:56 AM UTC

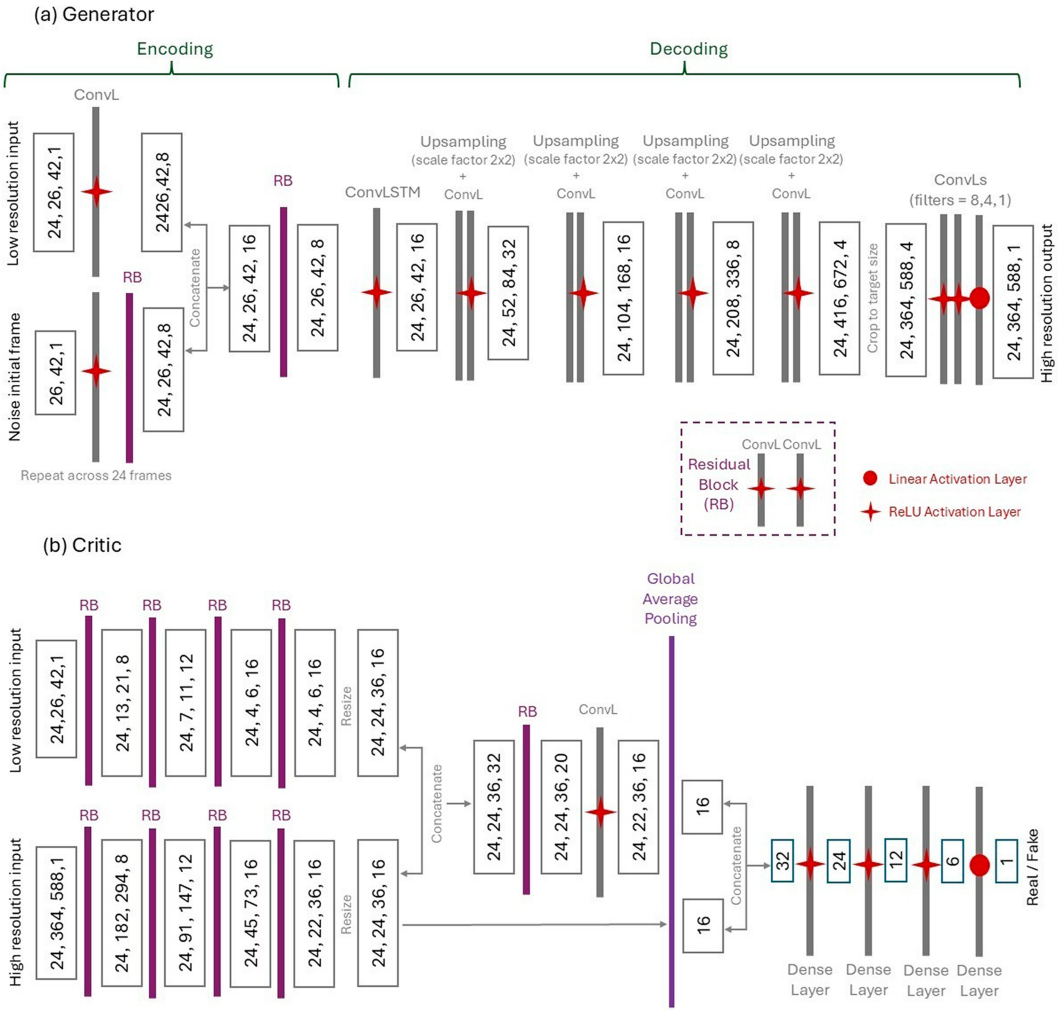


FIG. 2. The structure of the generator and critic networks used for the WGAN models. Each box represents the tensor dimensions as data pass through the network, expressed as (time steps, number of latitudes, number of longitudes, and number of channels). Convolutional layers use 3×3 kernels with ReLU activation unless otherwise stated; residual blocks comprise two such layers. Upsampling in the generator is performed at a scale factor of 2, followed by convolution, and the generator output uses a linear activation. Single numbers indicate the number of channels after flattening the remaining dimensions.

not negatively impact model performance (Leinonen et al. 2021).

The generator in the WGAN model is implemented as a U-Net architecture, defined by its encoder–decoder framework with skip connections that facilitate the learning of spatial patterns in the precipitation fields. The encoder compresses the input by reducing its spatial resolution while capturing multiscale features, and the decoder upsamples this representation to reconstruct high-resolution outputs. The generator network integrates a ConvLSTM layer that allows the model to map low-resolution precipitation inputs to high-resolution outputs while capturing temporal dependencies. The generator also includes a thresholding layer, ensuring that generated precipitation fields do not include negative values and effectively distinguishing between zero and nonzero precipitation regions.

The critic network begins by processing the low-resolution and high-resolution inputs through two separate branches, each containing multiple residual blocks. These branches are then merged, and the combined representation is passed through an additional residual block and convolutional layers. This design enables the critic to extract and integrate spatial features from both input resolutions effectively. While the critic does not include explicit temporal layers, it handles the temporal dimension using fully connected layers that flatten and process the third (time) dimension, allowing it to evaluate the realism of generated sequences.

b. Model training

WRF simulations for the historical and future periods each include 5386 days of hourly precipitation fields, from January 2001 to September 2015 and from January 2086 to September 2100,

Unauthenticated | Downloaded 10/09/26 09:56 AM UTC

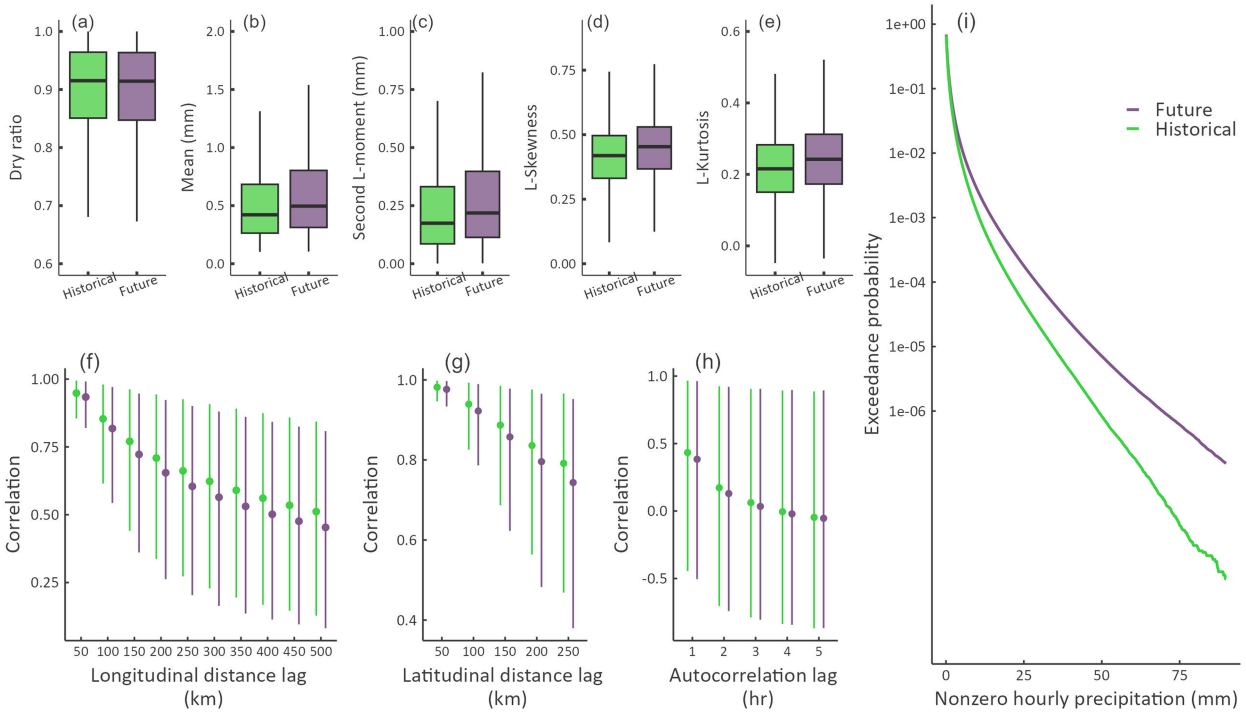


FIG. 3. Key statistics of historical and future 15-yr WRF precipitation blocks. (a) Dry ratio (fraction of dry grid cells per hourly field), (b)–(e) hourly L-moments, spatial correlations along (f) longitudinal and (g) latitudinal lags, (h) temporal autocorrelation up to 5 h, and (i) exceedance probability of pooled hourly nonzero precipitation. Boxplots show medians, interquartile ranges, and ±1.5 IQR whiskers; error bars show medians (points) with 95% confidence intervals.

respectively. Given the relatively short duration of these 15-yr datasets, we allocated the first 4846 days (approximately 90%) for training the deep learning models and reserved the remaining 540 days (approximately 10%) for evaluation. The historical model was therefore trained on the first 90% of the historical period and the future model on the first 90% of the future period. Evaluation was consistently performed using the final 540 days from each respective period, meaning that the same subsets (last 540 historical days and last 540 future days) were used across all experiments to ensure comparability of evaluation metrics.

Given the limited temporal extent of the datasets and the inherent variability of hourly precipitation, formal statistical tests for nonstationarity [e.g., Kwiatkowski–Phillips–Schmidt–Shin (KPSS) or Mann–Kendall test] may lack robustness and power in this context. Consequently, direct estimation of climate change signals (e.g., Δ future–historical) from these short records is also subject to considerable uncertainty. Instead, we treat the historical and future periods as distinct, stationary blocks and assess changes in distributional characteristics between them. This approach is further justified by the fact that each period is relatively short, while the temporal gap between them is sufficiently long, allowing nonstationarity to be inferred by directly comparing the historical and future distributions. The observed shifts in the precipitation distributions, particularly the emergence of heavier tails in the future period, suggest a departure from stationarity and an evolution through time of the underlying precipitation-generating processes (see section 4a, Fig. 3).

The U-Net model was not used as a stand-alone model but rather to preinitialize the generator of the WGAN. This strategic initialization provided the generator with physically meaningful weights, which improved convergence speed and reduced training time. The key advantage of the WGAN framework lies in its use of the Wasserstein loss, which estimates the divergence between the real and generated data distributions to better capture variability and extremes. By utilizing two competing networks, the generator and critic, trained in an adversarial manner, the WGAN is more complex to train and more sensitive to hyperparameters than U-Nets. To enhance stability, we adopted the U-Net-based initialization strategy, in which the generator weights were initialized using a pretrained U-Net model trained to reproduce high-resolution precipitation fields from low-resolution inputs. This initialization provides the generator with a physically meaningful starting point, reducing early stage instability and accelerating convergence, as demonstrated in prior studies (Abdelmoaty et al. 2025a; Frégier and Gouray 2021; Ham et al. 2020).

For the WGAN models, training was guided by optimizing the generator and critic using the Wasserstein loss with gradient penalty, as proposed by Gulrajani et al. (2017). This approach improves upon the original Wasserstein loss by replacing weight clipping, which was previously used to enforce the Lipschitz constraint but often led to limited model capacity and unstable training. The gradient penalty provides a more effective solution by encouraging the critic's gradients to maintain a norm close to one, thereby ensuring compliance with the Lipschitz condition.

Unauthenticated | Downloaded 10/09/26 09:56 AM UTC

This modification significantly enhances training stability and convergence (Leinonen et al. 2021). The training parameters were selected based on an iterative tuning process informed by established practices in the GAN literature (e.g., Bhatia and Dahyot 2019; Gulrajani et al. 2017; Leinonen et al. 2021; Radford et al. 2016; Tirel et al. 2024). The models were coded using TensorFlow–Python and trained on a high-performance computing cluster with graphics processing unit (GPU) acceleration. The training parameters are listed in Table S1 in the online supplemental material.

c. *Evaluation analysis*

We independently trained two models: one using historical simulations and the other using future simulations. To account for stochastic variability in the training process (e.g., random noise inputs, weight updates, and batch shuffling), each model was trained 10 times independently, starting from the same U-Net-initialized weights. The training performance followed the expected WGAN pattern with a gradient penalty (Fig. S1), with a steadily decreasing critic loss and a gradient penalty term that stabilized over epochs (Gulrajani et al. 2017; Petzka et al. 2018), with only a couple of runs exhibiting rising penalties, which were removed from further analysis steps. Each of these 10 realizations was then evaluated against both historical and future data held out for evaluation. Therefore, we have four evaluation scenarios: 1) hh—trained and tested on historical data, 2) hf—trained on historical data and tested on future data, 3) ff—trained and tested on future data, and 4) fh—trained on future data and tested on historical data. Each of the 10 model realizations was assessed across these four scenarios to evaluate their ability to reproduce key precipitation characteristics. Field-based characteristics included wet/dry occurrence probabilities, L-moments estimated for each hourly precipitation field, and spatial dependence across grid cells, while time series–based characteristics included temporal dependence, distributional properties, and extremes.

A threshold of 0.1 mm h^{-1} was applied to the hourly precipitation data to filter out trace amounts that fall below the detection capabilities of standard rain gauges and model resolution. This approach ensures that the quantification of loss metrics and other evaluation statistics focuses on meaningful precipitation events, minimizing distortion from spurious light rainfall. The 0.1 mm h^{-1} cutoff is widely adopted in observational and modeling studies to define "wet hours" and enhance statistical robustness (e.g., Lauer et al. 2023; Lei et al. 2022; Thomassen et al. 2022). Values below this threshold were set to zero, and thus all other statistics were derived using only the nonzero precipitation data (i.e., values exceeding 0.1 mm).

1) WET/DRY OCCURRENCES

To assess the ability of the downscaled fields to replicate the occurrence of wet and dry conditions, we evaluate the ratio of dry cells [R_0 as shown in Eq. (1)], defined as the proportion of grid cells with precipitation below the wet/dry threshold (0.1 mm h^{-1}) in each hourly field across the study domain. To quantify the deviation of the R_0 in the downscaled fields from the original ones, we calculated the bias as the average difference between the two fields:

$$R_0 = \frac{\text{number of cells with } P < 0.1}{\text{total number of grid cells}}, \tag{1}$$

where P is the precipitation.

2) L-MOMENTS

For each hourly precipitation field, we computed the mean, second L-moment, L-skewness, and L-kurtosis for both the downscaled and reference fields, reflecting the spatial distribution across grid cells with nonzero precipitation. L-moments were chosen over traditional moments due to their superior robustness and reliability in characterizing the statistical properties of precipitation, particularly in the presence of skewed or heavy-tailed distributions (Hosking 1990). Similarly, we calculated the bias as the average difference between the corresponding statistical measures of the downscaled and original fields.

3) SPATIAL AND TEMPORAL DEPENDENCE

The Pearson correlation coefficient [r; Eq. (2)] was computed between all grid values of the hourly precipitation field and the spatially shifted version of itself, in both the longitudinal (up to 10 gridcell lags, approximately 500 km) and latitudinal (up to 5 lags, approximately 250 km) directions. A shifted field refers to the original field displaced spatially by a specified number of grid cells in a particular direction. This approach is conceptually similar to temporal autocorrelation, where a value at a certain time is compared with its counterpart at a later time but applied here in the spatial domain. Spatial correlations are computed using all grid values in the precipitation field, including zero precipitation to evaluate the spatial coherence of the full precipitation field, including dry regions, which play an important role in the spatial organization of hourly precipitation. The spatial correlation diagnostic reflects the degree to which the downscaled fields preserve the spatial coherence and large-scale structural organization of precipitation in the reference dataset. To assess small-scale spatial variability, we additionally computed a high-wavenumber spectral power ratio, defined as the ratio of radially averaged power spectra at high spatial wavenumbers between the downscaled and reference fields [$R_{\text{high}-k}$; Eq. (3)]. Values of this ratio close to unity indicate that the downscaled fields retain a similar amount of small-scale variance as the reference, while values below unity indicate attenuation of fine-scale variability (spatial smoothing). Values above unity indicate excess small-scale variance, which may reflect unstructured or noisy fine-scale features rather than physically coherent structure (Harris et al. 2022; Iotti et al. 2025). For the overall similarity between the downscaled and reference precipitation fields, we also estimated the structural similarity measure. To complement the scale-dependent diagnostics, we also estimated the structural similarity index measure (SSIM) as a measure of overall similarity between the downscaled and reference precipitation fields (Tie et al. 2024).

Unauthenticated | Downloaded 10/09/26 09:56 AM UTC

In addition, traditional autocorrelation coefficients were calculated for both the downscaled and original time series to assess temporal dependence. Again, we estimated the bias to evaluate the deviation between the correlations of the downscaled and original fields:

$$r = \frac{\sum_{i=1}^{N}(X_i - \overline{X})(O_i - \overline{O})}{\sqrt{\sum_{i=1}^{N}(X_i - \overline{X})^2 \sum_{i=1}^{N}(O_i - \overline{O})^2}}, \quad (2)$$

where N is the number of grids, X_i is the shifted field value, O_i is the original field value, $\overline{X}$ is the average of the shifted field values, and $\overline{O}$ is the average of the original field values:

$$R_{\text{high}-k} = \frac{\sum_{k\in K_{\text{high}}} P_{\text{ds}}(k)}{\sum_{k\in K_{\text{high}}} P_{\text{ref}}(k)}, \quad (3)$$

where $P_{\text{ds}}(k)$ and $P_{\text{ref}}(k)$ denote the radially averaged two-dimensional power spectrum obtained from the fast Fourier transform for the downscaled and the reference datasets, respectively; and K_{high} represents the upper 30% of resolved spatial wavenumbers corresponding to the smallest spatial scales.

4) Distributional properties

First, we computed the Hellinger distance (H) to quantify the overall similarity between the probability density functions (PDFs) of the downscaled and original fields during the testing period (Hellinger 1909). The H distance is a well-established and robust method for comparing two PDFs [Eq. (4)]. The squared H distance takes values from 0 to 1, with 0 indicating a perfect match. Second, the Kolmogorov–Smirnov (KS) test statistic [Eq. (5)] was used to evaluate the similarity between the empirical cumulative distribution functions (CDFs) of the nonzero hourly time series from the simulated and reference fields during the testing period. The KS statistic ranges from 0 to 1, with 0 indicating a perfect match. PDFs were estimated using normalized histograms with fixed bin widths, and CDFs were obtained as the cumulative sums of these histograms. This approach provides an empirical, histogram-based approximation of the underlying distributions, computationally efficient for large datasets (~11 GB for each downscaled generation of hourly precipitation corresponding to the evaluation period):

$$H^2(p, q) = \frac{1}{2}\int_{x\in\mathcal{X}}\left[\sqrt{p(x)} - \sqrt{q(x)}\right]^2, \quad (4)$$

$$\text{KS} = \sup_x |F_X(x) - F_O(x)|, \quad (5)$$

where $p(x)$ and $q(x)$ are the two PDFs of the original and simulated data, respectively; $F_X(x)$ is the simulated PDF; $F_O(x)$ is the original PDF; and $\sup_x$ is the supremum (maximum) over all values of x.

5) Extremes

To evaluate the model's ability to capture extreme precipitation behavior, the upper tail of the distribution was analyzed using the 75th, 95th, 99th, and 99.9th percentiles of hourly nonzero precipitation during the testing period, as well as the daily maxima. Percentiles are computed from the pooled distribution of nonzero hourly precipitation values across all grid cells and the entire test period (i.e., a space–time marginal distribution). In contrast, daily maxima are obtained by computing the maximum hourly precipitation within each 24-h period at each grid cell. By comparing these extreme indices for the downscaled and original fields, we assess the models' skill in reproducing moderate to extreme precipitation events. This approach provides a straightforward and distribution-independent evaluation of extremes, capturing changes in intensity and frequency without assuming a specific parametric form. Biases were computed as the mean difference between the percentiles of the downscaled and original fields, highlighting any over- or underestimation of extreme values across varying intensities.

4. Results

a. *Comparison between historical and future WRF simulations*

Before evaluating the downscaled fields from the different experiments, we first examine the raw WRF outputs to highlight the shift toward nonstationarity in the future simulations relative to the historical period. The future simulations exhibit a slight increase in the magnitude of key distributional characteristics, including the mean, second L-moment, L-skewness, and L-kurtosis (Figs. 3b–e). These results suggest a shift toward higher, more variable, and more asymmetric precipitation distributions with heavier tails. In contrast, the probability of dry cells (Fig. 3a) remains largely unchanged between the historical and future periods. This lack of change in dry spell frequency may reflect limitations of the modeling framework in capturing potential shifts in wet/dry spell dynamics under future climate conditions (Abdelmoaty et al. 2025b; Tabari et al. 2021; Vrac et al. 2007). Figures 3f and 3g show a slightly faster decay of spatial correlation under future conditions, both longitudinally and latitudinally, accompanied by larger spread across realizations. This indicates a modest reduction in spatial coherence of precipitation fields, consistent with a tendency toward smaller or less organized storm cells in a warmer climate. However, temporal autocorrelation (Fig. 3h) remains relatively consistent between the two periods across multiple lags, suggesting that the persistence of precipitation events over time may not substantially change. The most prominent distinction appears in the extremes, as highlighted by the exceedance probability curves (Fig. 3i). The future distribution shows clear dominance in the upper tail, indicating more frequent extreme hourly precipitation events, and is perhaps the most prominent mark of nonstationarity.

b. *Assessment of wet/dry occurrence*

Turning to the evaluation of the WGAN downscaled outputs, we first assess wet/dry occurrences within the hourly fields. When the model is trained and tested within the same climate period (i.e., hh and ff cases), it tends to reproduce the proportion of dry grid cells with relatively low bias proportion

Unauthenticated | Downloaded 10/09/26 09:56 AM UTC

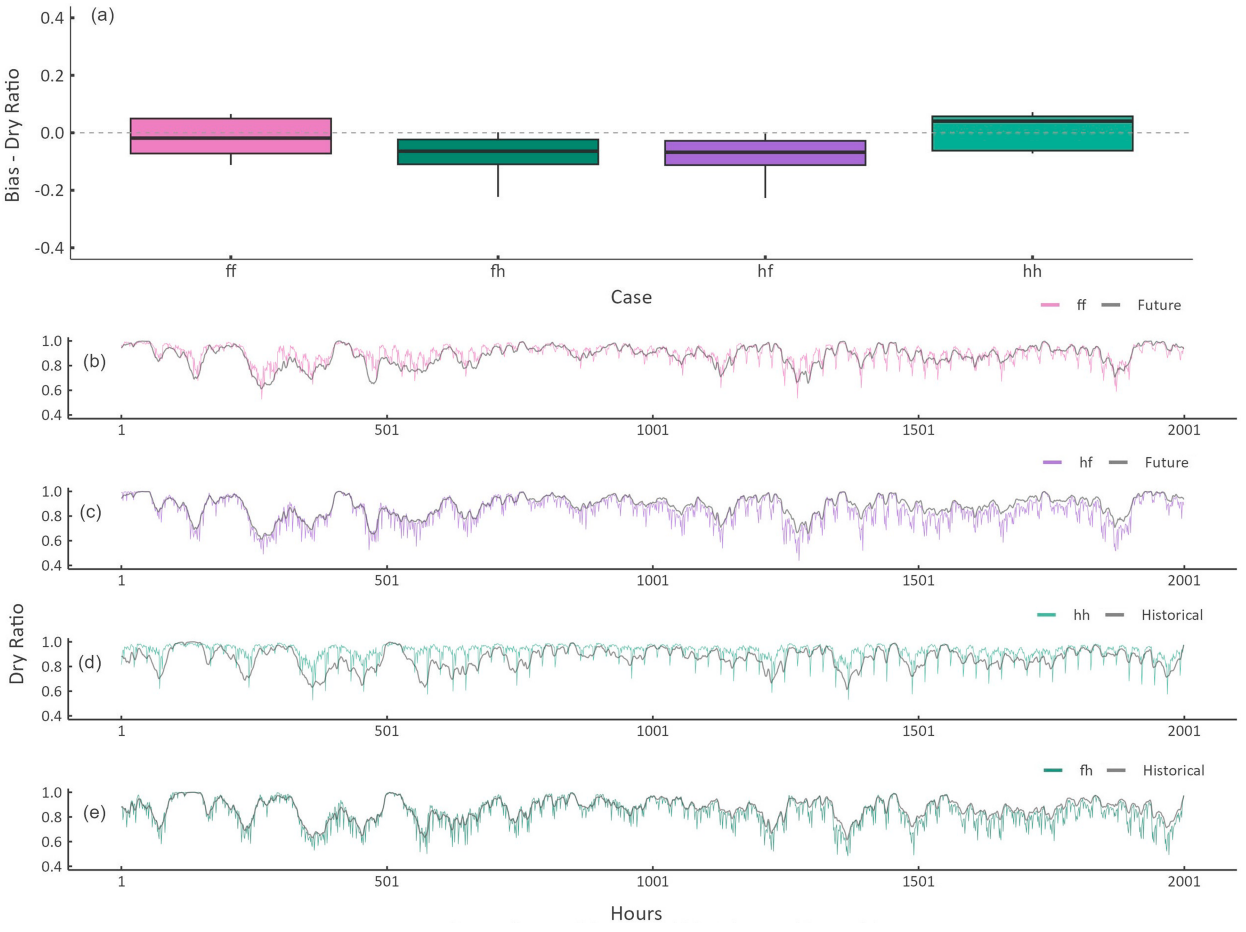


FIG. 4. Evaluation of wet/dry grid downscaling. (a) Bias in dry ratio (fraction of grid cells ≤ 0.1 mm h^{-1}) between original and downscaled fields; negative values indicate a dry bias, and positive values indicate a wet bias. Boxplots show medians, interquartile ranges, and ± 1.5 IQR whiskers across all hourly fields and 10 realizations. Time series of dry ratio for each case showing the first 2000 time steps: (b) ff (future–future), (c) hf (historical–future), (d) hh (historical–historical), and (e) fh (future–historical), averaged over 10 realizations, with original values shown for comparison.

(medians of 0.040 and −0.018, respectively; Fig. 4a) and good agreement with the reference probability dry time series (Figs. 4b,d). This consistency indicates the model's ability to learn and replicate the wet/dry occurrence patterns under the same climatic conditions. In contrast, models evaluated under a different climate regime than the one used for training (fh and hf cases) show a slight underestimation of the dry ratio (with proportion bias medians of −0.064 and −0.068, respectively), suggesting challenges in generalizing across shifting climate conditions for dry/wet conditions. This performance is also reflected in their respective time series (Figs. 4c,e), which closely follow the fluctuations of the original dry probabilities, though with pronounced troughs at certain time steps that lead to underestimation. Overall, these results highlight the model's reasonable ability to reproduce wet/dry characteristics across all evaluation setups, with particularly strong performance of the within-climate cases (hh and ff) and notable generalization capacity in cross-climate cases (hf and fh), despite minor biases, demonstrating the framework's adaptability to both historical and future climate conditions.

c. *Assessment of L-moments*

When evaluating the L-moments, the fh and hf cases, where models are tested on a different climate period than they were trained on, demonstrate the least spread in bias for both the mean and second L-moment (Figs. 5a,b). The interquartile range (IQR) of biases spans from −0.15 to −0.04 for the fh case and from −0.23 to −0.11 for the hf case, compared to −0.14 to 0.08 and −0.27 to −0.05 for the hh and ff cases, respectively. This result highlights that cross-climate cases consistently stay within a bias range of less than 0.2. This narrower spread in the cross-climate cases (hf and fh) likely reflects a smoothing effect, while the within-climate models (hh and ff) produce more heterogeneous patterns (Figs. 5e,f). Examples are shown in Figs. 5e and 5f, and more examples are shown in Figs. S2 and S3. Overall, all four cases exhibit small but consistently negative biases, indicating that the downscaled fields tend to slightly underestimate the central tendency and variability of hourly precipitation compared to the original fields. Biases in higher-order moments (Figs. 5c,d) are also minimal: L-skewness interquartile ranges

Unauthenticated | Downloaded 10/09/26 09:56 AM UTC

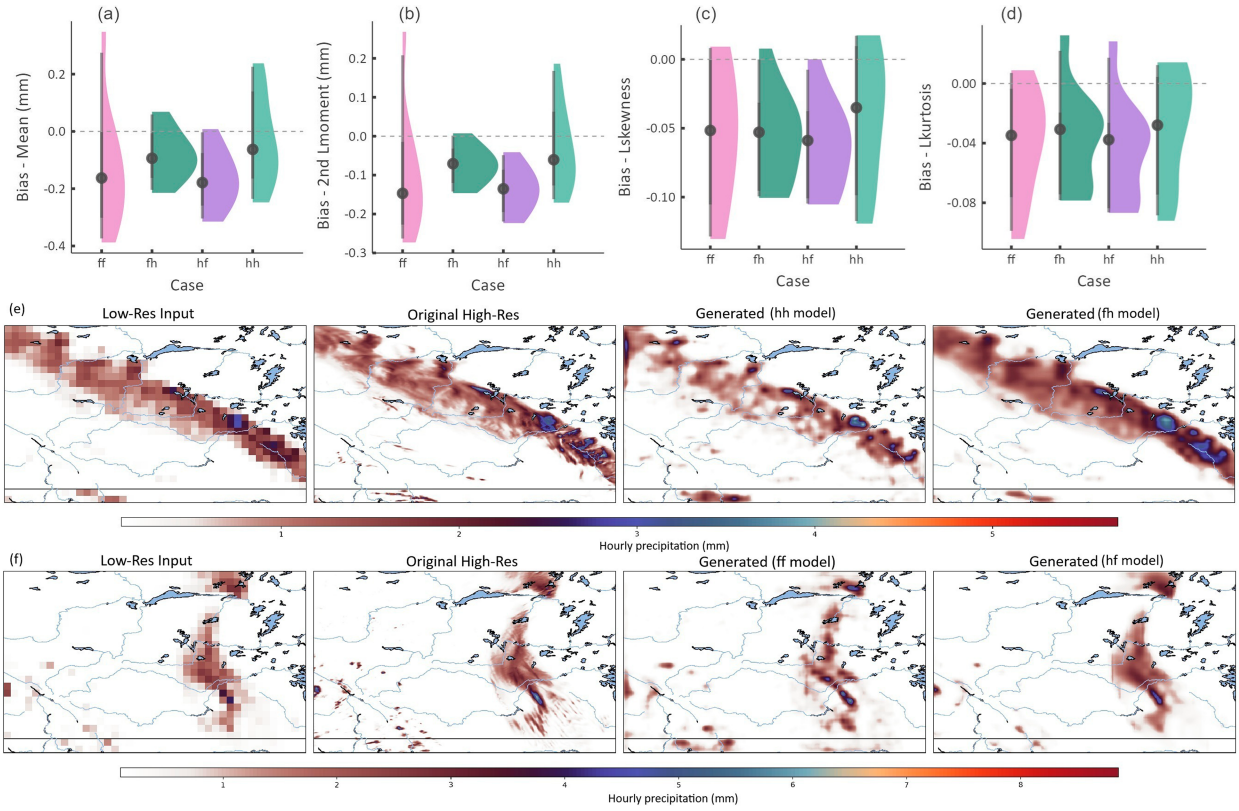


FIG. 5. Model performance in reproducing L-moments. Biases of hourly (a) mean, (b) second L-moment, (c) L-skewness, and (d) L-kurtosis. Shaded regions and error bars indicate the spread across all hourly fields and 10 realizations, with medians shown as points. Results are shown for four cases: ff (future–future), fh (future–historical), hf (historical–future), and hh (historical–historical). Example hourly fields: (e) historical simulations predicted by hh and fh models, and (f) future simulations predicted by ff and hf models.

span roughly from −0.09 to 0.00, and L-kurtosis ranges from approximately −0.08 to 0.00 across all hourly fields and the 10 realizations for the four cases, corresponding to changes within about ±0.05 from the median. Given that typical L-skewness and L-kurtosis values for precipitation are in the range of ~0.2–0.3 and ~0.1–0.25, respectively (Sankarasubramanian and Srinivasan 1999), these biases represent relatively minor deviations. Collectively, these results indicate that the models are broadly effective in reproducing not only the central tendency but also the shape of the precipitation distribution across the field, maintaining reliable performance even when applied across different climate periods.

d. Assessment of temporal and spatial autocorrelations

Models trained and tested within the same climate period (hh and ff) consistently exhibit the lowest biases in reproducing spatial autocorrelations, with values remaining below 0.22 for the longitudinal and 0.1 for the latitudinal directions across all lags (Figs. 6a,b). Notably, in these cases, longitudinal correlation biases peak at 0.16 and 0.22 for the hh and ff models, respectively, at distances of 150–200 km. In contrast, cross-climate applications (hf and fh) display noticeably larger positive biases, particularly at medium to long distances: Longitudinal correlation biases peak at 0.32 for hf and 0.27 for fh around 200–300 km before gradually declining, while latitudinal biases increase almost linearly with distance, reaching ~0.2 at 250 km for hf compared to <0.1 for hh and ff. While hh and ff configurations exhibit better reproduction of the spatial correlation, spectral diagnostics reveal a modest attenuation of small-scale power relative to the reference. In contrast, cross-climate configurations (hf and fh) sometimes display comparable or higher high-wavenumber power, reflecting increased small-scale variance (Fig. S5a). However, this excess power does not necessarily correspond to physically coherent fine-scale structure and is often associated with noisier textures. SSIM values are consistently high across all configurations, indicating strong overall structural similarity between the downscaled and reference precipitation fields (Fig. S5b). Together, these results suggest that within-climate training better preserves coherent spatial organization, while cross-climate application may introduce unstructured small-scale variability. For temporal autocorrelation (Fig. 6c), all cases exhibit similar patterns: an initial underestimation at short lags (from −0.25 to −0.08 at 1 h) followed by convergence toward zero bias beyond ~4–5 h.

e. Assessment of the overall distribution

Quantifying the distributional similarity using the H distance and KS statistic, the hh case exhibits the lowest values for both metrics (median $H = 0.079$, KS = 0.077) with relatively narrow variability (Figs. 7c,d), indicating strong agreement between

Unauthenticated | Downloaded 10/09/26 09:56 AM UTC

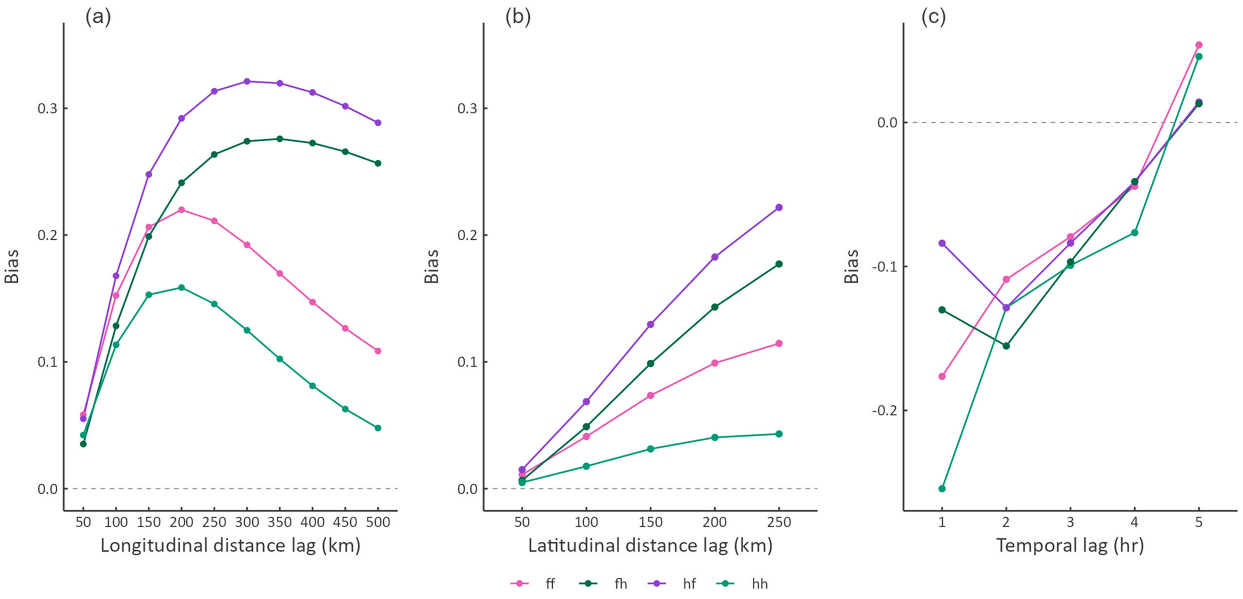


FIG. 6. Biases in Pearson correlation coefficients for (a) longitudinal, (b) latitudinal, and (c) temporal (autocorrelation) correlations across the four cases: ff, fh, hf, and hh. Bias at each lag is computed as the mean difference between pooled correlations from 10 realizations and the corresponding observed values. Full correlation structures are shown in Fig. S4.

the downscaled and original distributions for the within-climate cases. The ff case shows slightly higher metrics (median $H = 0.092$, KS = 0.092) but maintains relatively stable performance, reflected by the low variability. In contrast, the cross-climate cases (hf and fh) display noticeably larger values (median $H = 0.134$ and $= 0.122$; KS $= 0.175$ and $= 0.160$, respectively) and wider spreads, reflecting increased uncertainty in training and prediction when applied to a different climate regime. Examination of the nonexceedance probability curves (Figs. 7a,b) shows that the same-period models (hh and ff) more closely track the original distributions across the low- to moderate-precipitation range, while the cross-climate cases deviate more noticeably, particularly in the low-precipitation range (<1 mm). This finding indicates that distributional mismatches in cross-climate models are more pronounced in the central portion of the distribution.

f. Assessment of extremes

All models systematically underestimate extreme precipitation values compared to the raw WRF simulations, with the negative bias becoming more pronounced toward the upper percentiles (Fig. 8a). Biases are minimal at the 75th percentile (from −0.47 to 0.25 mm across cases) but deepen substantially at higher quantiles, reaching approximately −4.3 mm at the 99th percentile and as much as −9.9 mm at the 99.9th percentile for models trained or evaluated on future data (ff and hf). In contrast, models evaluated on past data (hh and fh) exhibit slightly smaller underestimations at these extremes (−2.7 and −7.1 mm at the 99th and 99.9th percentiles, respectively), reflecting both the generally less intense historical extremes and a better model fit to those conditions. Notably, the bias gap between future- and past-period evaluations grows with percentile level, indicating that nonstationarity amplifies model limitations at the most extreme events. Although training on future data (as in ff) reduces bias compared to hf at some quantiles, it remains insufficient to fully correct underestimations at the rarest events. The daily maxima show a similar performance pattern: best performance with the hh model, followed by fh, ff, and hf, respectively (Figs. 8c,d). It was expected that models trained on the future and tested on the past would reproduce historical extremes; however, the model trained and tested on historical data showed the best performance. Models tested on future data (ff and hf) had slightly higher (more negative) biases, consistent with the more heavy-tailed future distributions. Overall, all models show small biases in reproducing the daily maxima, ranging from −0.21 to −0.57 mm.

5. Discussion

The WRF Model represents one of the most advanced developments in convection-permitting modeling, offering some of the finest spatial resolutions currently available for regional climate simulations (Prein et al. 2015). The WRF Model dataset from Li et al. (2019) used in this study exhibits nonstationarity in the PGW simulations for the future period, arising from the imposed climate change perturbations. In their setup, future boundary conditions were generated by adding monthly CMIP5-derived change signals to ERA-Interim reanalysis, introducing both seasonally varying mean shifts and altered vertical moisture and temperature profiles. This modification not only changes the central tendency of key meteorological variables but also reshapes their probability distributions (Fig. 3), reflecting a departure from stationarity. Similar behavior has been reported in other PGW studies; for instance, Lenderink et al. (2025) reported that extreme rainfall in convection-permitting PGW experiments intensified beyond Clausius–Clapeyron

Unauthenticated | Downloaded 10/09/26 09:56 AM UTC

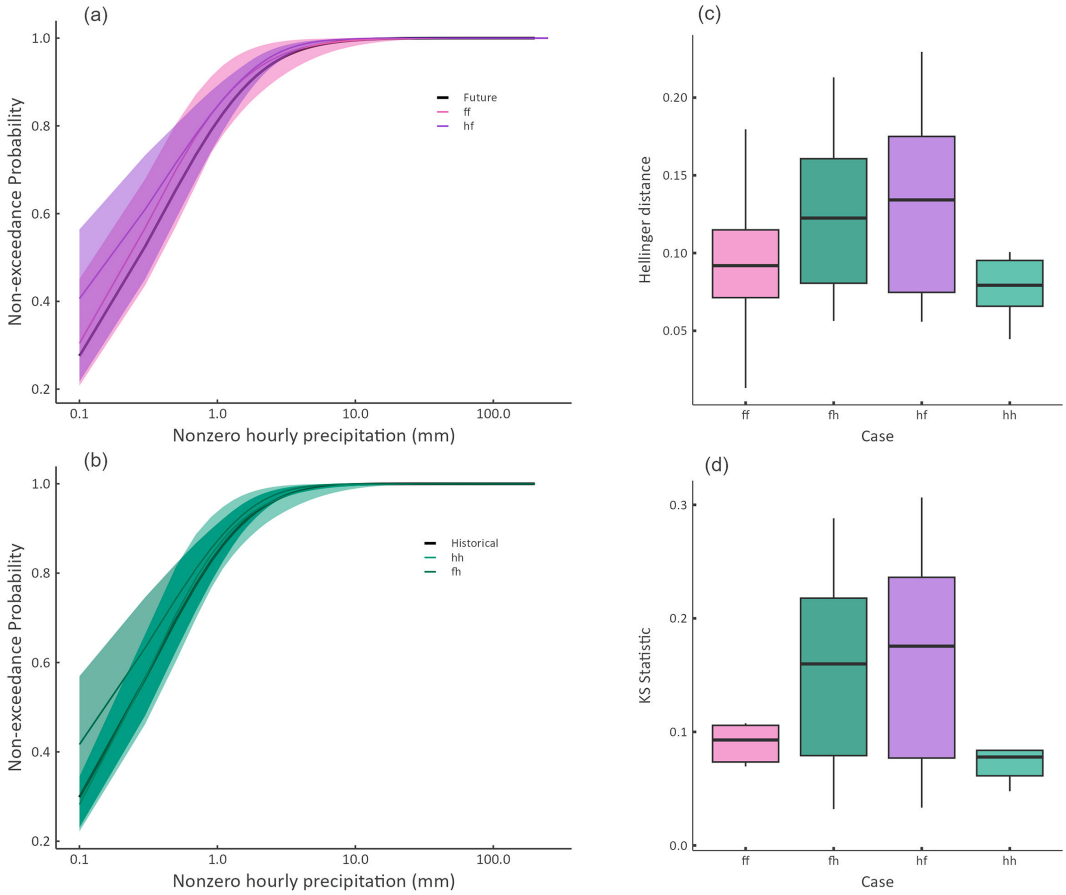


FIG. 7. Evaluation of distributional similarity. (a) Nonexceedance probability for future data compared with downscaled outputs from models trained on future (ff) and past (hf). (b) As in (a), but for historical data compared with hh and fh. (c) Hellinger distance and (d) KS statistic quantify differences between original and downscaled distributions across the four cases. For probability plots, lines represent the average of 10 realizations, and shaded regions indicate their variability. Boxplots show medians and variability across realizations.

expectations, accompanied by spatial shifts in heavy precipitation patterns. Similarly, Tahara et al. (2025) found out that PGW in the WRF Model intensifies precipitation at high resolution under various warming signals, highlighting the nonstationarity shifts in extreme precipitation. However, an inherent limitation of the PGW framework, as highlighted by Li et al. (2019), is that it does not account for changes in circulation patterns or shifts in storm-track frequency, which could further influence the manifestation of nonstationarity in the future climate. Therefore, while the dataset confirms significant statistical shifts in precipitation characteristics under PGW forcing, the true extent of nonstationarity may be even greater when considering dynamic changes not captured by the method.

The proportion of dry hours in the WRF Model simulations (Fig. 3a) remains essentially unchanged between the historical and future periods; however, this does not imply stationarity in wet/dry dynamics. Li et al. (2019) show that future precipitation shifts toward fewer light events and more intense extremes, with longer interevent intervals in summer, meaning that dry-hour fractions can remain stable even as intensity and sequencing change. Similar behavior has been documented in previous studies, where warming leads to heavier events separated by longer dry spells (Ficklin et al. 2022; Trenberth et al. 2003), and regional studies show that aggregation across seasons or basins can mask significant increases in summer dry-spell length (Zolina et al. 2013). The apparent stability in dry-hour probability may also stem from limitations of the PGW framework and model configuration in representing shifts in wet/dry spell dynamics under future climate conditions (Abdelmoaty et al. 2025b; Tabari et al. 2021; Vrac et al. 2007), as well as from threshold sensitivity, trace precipitation biases that inflate wet-hour counts in both periods (Maraun 2013; Piani et al. 2010), and PGW's omission of large-scale circulation or storm-track changes (Rasmussen et al. 2011). Thus, while this metric appears unchanged, the underlying precipitation occurrence processes could still be nonstationary.

Across the four cases of training the models using historical and future simulations and testing them within- and across-climate periods, precipitation statistics show a slight underestimation (Fig. 5), primarily for events likely driven by small, localized, intense storms that occupy only a single or a few 4-km grid cells. While the 50–4-km downscaling effectively captures

Unauthenticated | Downloaded 10/09/26 09:56 AM UTC

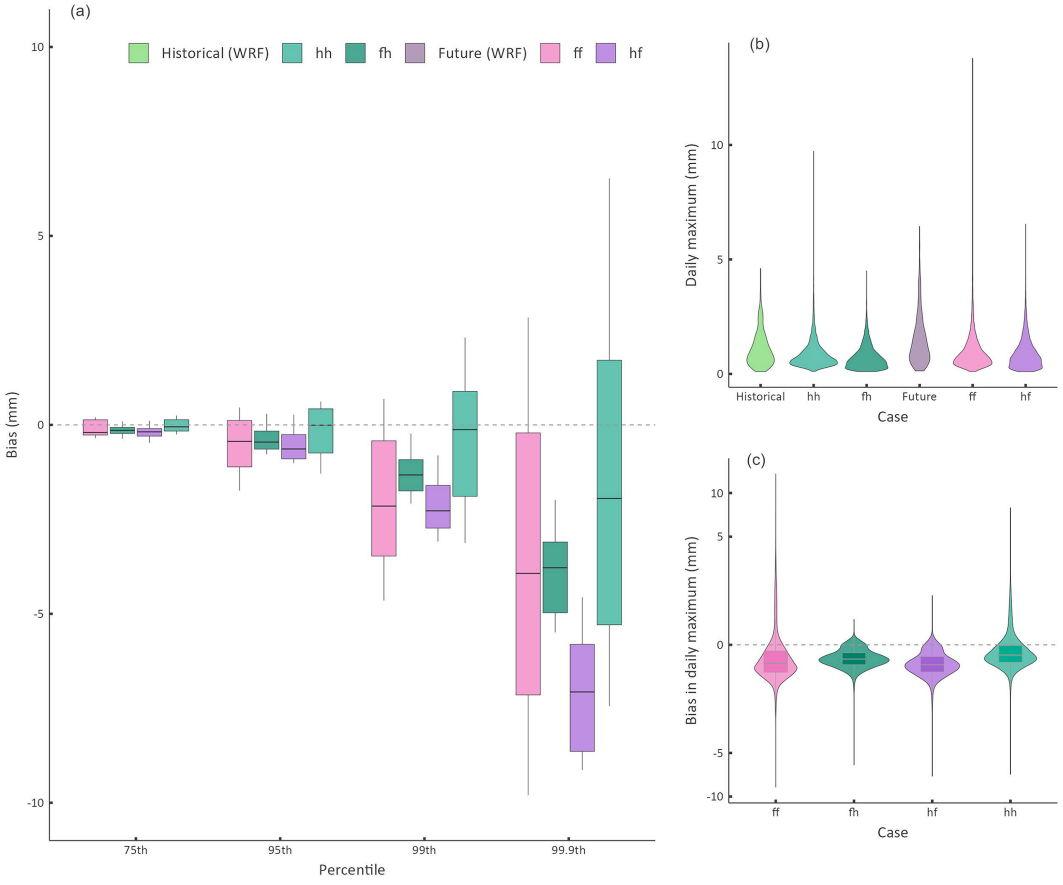


FIG. 8. Evaluation of extremes. (a) Biases in the 75th, 95th, 99th, and 99.9th percentiles computed from pooled nonzero hourly precipitation values across all grid cells; boxplots show medians and variability across 10 realizations per case. (b) Daily maxima, defined as the maximum hourly precipitation within each 24-h period at each grid cell, and (c) the boxplot and violin plot of biases of daily maxima across the four cases: ff (future–future), fh (future–historical), hf (historical–future), and hh (historical–historical). Violin plots show variability across all grid cells and days in the evaluation period and across realizations.

larger organized systems that span multiple grid cells, the most localized convective bursts remain more difficult to reproduce. These small-scale events have sharp spatial gradients and short lifespans, making them challenging to learn from the available training data, especially if their representation is sparse in the historical record. In our WGAN-based downscaling, small, highly localized convective bursts tend to be slightly underestimated, likely because deep learning architectures often prioritize broader-scale structural consistency over isolated high-intensity peaks (Reddy and Matear 2025; Wang et al. 2021). In our case, the underestimation is minor, indicating that the WGANs reproduce the overall precipitation distribution well, but smoothing at the upper tail may occur for the most localized extremes. Targeted data augmentation, specialized loss components focusing on localized intensity features, or stepwise downscaling could further improve performance for these rare, small-footprint events.

The within-climate WGAN models (hh and ff) appear more skillful at reproducing the heterogeneous spatial patterns of precipitation, resulting in slightly noisier downscaled fields compared to the smoother outputs from the cross-climate models (hf and fh) as shown in Fig. 5. This increased spatial variability likely reflects a closer alignment with the fine-scale heterogeneity present in the target WRF Model precipitation fields, which can enhance spatial correlation performance (Vandal et al. 2019). Indeed, the better performance of hh and ff models in reproducing spatial correlations (Figs. 6a,b) suggests that they are better able to replicate the heterogeneous spatial distribution of precipitation, potentially because the spatial correlation structure is more consistent within the same climate period and diverges slightly between historical and future WRF Model simulations.

Because the spatial correlation metric primarily reflects large-scale spatial structure and coherence, it may overlook small, isolated storm features, whereas the radial power spectrum at high wavenumbers is designed to capture small-scale spatial variability. In this sense, the two diagnostics provide a complementary overview of the spatial organization. Consistent with this interpretation, while the within-climate models slightly outperform in capturing large-scale spatial organization, the cross-climate

Unauthenticated | Downloaded 10/09/26 09:56 AM UTC

models tend to preserve small-scale variance somewhat better, although all four model configurations remain close to unity, indicating relatively modest differences (Fig. S5).

By contrast, cross-climate models may be smoothing fields more aggressively due to a mismatch between the spatial statistics learned during training and those required for prediction, which is a reasonable outcome given that within-climate models focus on learning the target distribution, whereas cross-climate models introduce additional smoothing of spatial features. This behavior has been noted in previous deep learning downscaling studies, where training–prediction climate differences degrade spatial structure reproduction (Baño-Medina 2020; Pan et al. 2019).

In contrast, the temporal autocorrelation is well reproduced in all cases (Fig. 6c), which is expected given that the autocorrelation structure of hourly precipitation appears relatively unchanged between the historical and future WRF Model simulations. This stability suggests that temporal persistence characteristics are less sensitive to nonstationarity than spatial patterns, aligning with previous findings in both physical and statistical downscaling frameworks (Cannon et al. 2015; Wang et al. 2021). Finally, the hh and ff models also reproduce the overall precipitation distributions more accurately (reflected in their lower H distance and KS statistics with less variability; Fig. 7) likely because the distributions are more similar within a climate period than across climates. This within-climate similarity reduces the distributional shift between training and prediction datasets, a factor known to improve statistical fidelity in generative and deep learning downscaling approaches (Baño-Medina 2020; Vandal et al. 2019).

Regarding extremes (Fig. 8a), the hh model (trained and evaluated on the historical simulation) shows the smallest biases across the 75th, 95th, 99th, and 99.9th percentiles, consistent with expectations that within-climate training/testing reduces distributional mismatch and yields more reliable reproduction of precipitation extremes. Biases increase with percentile across all models, reflecting the statistical difficulty of capturing rare events from a limited sample size. The fh model (trained on future, evaluated on historical) ranks second, likely because exposure to a more extreme future climate during training improved its ability to capture high-end quantiles in the milder historical climate; similar results are obtained when testing daily precipitation downscaled by a GAN-based model over New Zealand (Rampal et al. 2024). The ff model also shows relatively low bias at higher percentiles but with considerably higher variability and uncertainty, which may stem from the more skewed and variable nature of the future precipitation distribution. By contrast, the hf model (trained on historical, evaluated on future) exhibits the largest biases, highlighting the challenges of applying models trained on less-skewed distributions to predict more variable and extreme future climates. The daily maximum precipitation analysis (Figs. 8b,c) further confirms these results. All four models exhibit low mean biases in daily maxima, indicating that they reproduce overall intensity distributions reasonably well. However, the ranking of performance mirrors the percentile-based analysis: hh performs best, followed by fh, with ff and hf showing progressively weaker skill. This pattern reflects the well-known dilemma of nonstationarity: Models trained within the same climate regime perform best, while cross-climate applications suffer from distribution shift and degraded tail performance (Baño-Medina 2020; Pan et al. 2019; Vandal et al. 2019).

These findings suggest that future applications of deep learning downscaling must explicitly account for nonstationarity in climate distributions. Training exclusively on historical data may underestimate the intensity and variability of future extremes, while training only on future data may inflate variance and uncertainty (Rasp and Thuerey 2021; Vandal et al. 2019). Hybrid or transfer-learning strategies that combine historical and future training samples could mitigate these issues by broadening the model's representation of both typical and extreme conditions (Baño-Medina 2020; Beucler et al. 2024). Moreover, given the increasing skewness of precipitation under warming (Papalexiou and Montanari 2019), extreme-aware loss functions or tail-sensitive generative methods may be necessary to capture upper-percentile behavior better (Reddy et al. 2023). Importantly, however, our results also show that the biases of cross-climate models were not substantially larger than those of within-climate models across the evaluation metrics. Even for extremes, the differences largely fell within the uncertainty range of hh and ff. This indicates that WGANs have promising potential, offering partial transferability of the downscaling relationship across climates. In practical terms, this challenge has direct consequences for infrastructure design and risk assessment, where underestimation of rare extremes can lead to significant underpreparedness (Allan and Soden 2008; Fowler et al. 2021). Thus, improving model robustness under distribution shifts remains a key research frontier for deep learning approaches in hydroclimate applications.

6. Conclusions

In this study, we examined the dilemma of training deep learning models for downscaling hourly precipitation under the assumption of stationarity and then applying them to nonstationary future conditions. Specifically, we developed and tested two WGAN models: one trained on historical data (2000–15) and another on future data (2086–2100), both derived from convection-permitting WRF simulations over western Canada. Our objective was to assess how sensitive the WGAN-based downscaling task is to nonstationarity by evaluating the models within and across climate periods. This resulted in four experimental cases: hh (trained and tested on historical data), ff (trained and tested on future data), fh (trained on future and tested on historical data), and hf (trained on historical and tested on future data). The key findings can be summarized as follows:

- Within-climate models (hh and ff) achieved stronger overall performance across most metrics, reproducing more heterogeneous and realistic spatial precipitation patterns, albeit with slightly higher variability in the mean and variance. However, the biases for all cases were comparable.
- Spatial autocorrelations in the longitudinal and latitudinal directions, as well as the probability of dry hours, were slightly better captured by within-climate models, reflecting their ability to represent the spatial heterogeneity of precipitation.

Unauthenticated | Downloaded 10/09/26 09:56 AM UTC

- The temporal autocorrelation structure was consistently reproduced across all four cases, supported by the relatively stable autocorrelation structure between historical and future climates.
- Regarding extremes, all four cases underestimated extremes in terms of percentile during the testing period and the daily maxima, though with important distinctions:
- Within-climate models (hh and ff) performed best, but the ff case exhibited high variability due to the skewness (heavy right tail) of the future data distribution.
- The fh model also performed reasonably well, likely benefiting from training on future data that contained stronger extremes.
- The hf model performed poorest for extremes, underscoring the difficulty of transferring models trained on milder historical climates to more variable and skewed future climates.

Our findings highlight the fundamental challenge of nonstationarity in climate downscaling: Models trained in one climate regime do not always generalize well to another, particularly when the distribution of precipitation becomes more skewed and heavy-tailed in the future. While within-climate training provides the best performance, this strategy is impractical for real-world applications where only historical observations are available. Importantly, however, our results also show that the cross-climate models did not perform substantially worse than the within-climate models, particularly for L-moments, dry probability, and temporal autocorrelation, and even for extremes, their differences generally fell within the uncertainty range of the hh and ff models. These findings suggest that WGANs offer a potential for at least partial transferability of the learned downscaling relationship across climates. However, this robustness is likely dependent on the degree of nonstationarity present: If the distributional shifts were much more pronounced than in our application, the performance of WGANs could deteriorate considerably. A controlled set of experiments, for example, using synthetic data with prescribed forms and magnitudes of nonstationarity, would be valuable to systematically assess the limits of model generalization. These insights underscore the need for approaches that explicitly address distribution shift, such as transfer learning, hybrid training with both historical and projected simulations, or climate-invariant representations that can generalize across changing regimes. In addition, methods that place greater emphasis on extremes, through tail-sensitive loss functions, stochastic downscaling components, or integration with extreme value theory, may help reduce underestimation at the highest percentiles.

Acknowledgments. H. M. A. is funded by the Mitacs Elevate Program in partnership with BGC Engineering Inc. S. M. P. is funded by the Natural Sciences and Engineering Research Council of Canada (NSERC Discovery Grant: RGPIN-2019-06894). A. M. and S. S. gratefully acknowledge the financial support provided by the Environmental Institute, University of Virginia. The authors are sincerely grateful to Professor John Pomeroy for sponsoring H. M. A.'s supercomputer account with the Digital Research Alliance of Canada, through the Global Water Futures Observatories Project at the Global Institute for Water Security. The authors gratefully acknowledge Dr. Yanping Li for facilitating access to the WRF simulations over western Canada. The authors declare no competing interests.

Data availability statement. The WRF simulations are publicly available through the following link: https://gwf.usask.ca/core-modelling/data-and-models/meteorological-forcing-data.php. The codes for this analysis are stored in GitHub through this link: https://github.com/HebaMSA/GAN-based-Downscaling02 (Abdelmoaty 2026).

REFERENCES

Abdelmoaty, H. M., 2026: GAN-based downscaling02 Version 1.0.0. Zenodo, accessed 28 December 2025, https://doi.org/10.5281/zenodo.18188397.

——, S. M. Papalexiou, A. Mamalakis, S. Singh, V. Coia, M. Hairabedian, P. Szeftel, and P. Grover, 2025a: Generative adversarial networks for downscaling hourly precipitation in the Canadian prairies. *J. Geophys. Res. Mach. Learn. Comput.*, **2**, e2025JH000678, https://doi.org/10.1029/2025JH000678.

——, C. R. Rajulapati, S. D. Nerantzaki, and S. M. Papalexiou, 2025b: Bias-corrected high-resolution temperature and precipitation projections for Canada. *Sci. Data*, **12**, 191, https://doi.org/10.1038/s41597-025-04396-z.

Albawi, S., T. A. Mohammed, and S. Al-Zawi, 2017: Understanding of a convolutional neural network. *2017 Int. Conf. on Engineering and Technology (ICET)*, Antalya, Turkey, Institute of Electrical and Electronics Engineers, 1–6, https://doi.org/10.1109/ICEngTechnol.2017.8308186.

Allan, R. P., and B. J. Soden, 2008: Atmospheric warming and the amplification of precipitation extremes. *Science*, **321**, 1481–1484, https://doi.org/10.1126/science.1160787.

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.

Balduzzi, D., M. Frean, L. Leary, J. P. Lewis, K. W.-D. Ma, and B. McWilliams, 2018: The shattered gradients problem: If resnets are the answer, then what is the question? arXiv, 1702.08591v2, https://doi.org/10.48550/arXiv.1702.08591.

Baño-Medina, J., 2020: Understanding deep learning decisions in statistical downscaling models. *CI2020: Proceedings of the 10th International Conference on Climate Informatics*, ACM, 79–85, https://doi.org/10.1145/3429309.3429321.

Beucler, T., and Coauthors, 2024: Climate-invariant machine learning. *Sci. Adv.*, **10**, eadj7250, https://doi.org/10.1126/sciadv.adj7250.

Bhatia, S., and R. Dahyot, 2019: Using WGAN for improving imbalanced classification performance. *CEUR Workshop Proc.*, Galway, Ireland, CEUR, 365–375, https://ceur-ws.org/Vol-2563/aics_34.pdf.

Bouvrie, J., 2006: Notes on convolutional neural networks. *Neural Nets*, **10**, 43–51.

Brooks, H. E., 2013: Severe thunderstorms and climate change. *Atmos. Res.*, **123**, 129–138, https://doi.org/10.1016/j.atmosres.2012.04.002.

Cannon, A. J., S. R. Sobie, and T. Q. Murdock, 2015: Bias correction of GCM precipitation by quantile mapping: How well do methods preserve changes in quantiles and extremes? *J. Climate*, **28**, 6938–6959, https://doi.org/10.1175/JCLI-D-14-00754.1.

Unauthenticated | Downloaded 10/09/26 09:56 AM UTC

Chan, S. C., E. J. Kendon, H. J. Fowler, S. Blenkinsop, C. A. T. Ferro, and D. B. Stephenson, 2013: Does increasing the spatial resolution of a regional climate model improve the simulated daily precipitation? *Climate Dyn.*, **41**, 1475–1495, https://doi.org/10.1007/s00382-012-1568-9.

Chaudhuri, C., and C. Robertson, 2020: CliGAN: A structurally sensitive convolutional neural network model for statistical downscaling of precipitation from multi-model ensembles. *Water*, **12**, 3353, https://doi.org/10.3390/w12123353.

Chen, D., A. Dai, and A. Hall, 2021: The convective-to-total precipitation ratio and the "drizzling" bias in climate models. *J. Geophys. Res. Atmos.*, **126**, e2020JD034198, https://doi.org/10.1029/2020JD034198.

Chen, H., J. Sun, W. Lin, and H. Xu, 2020: Comparison of CMIP6 and CMIP5 models in simulating climate extremes. *Sci. Bull.*, **65**, 1415–1418, https://doi.org/10.1016/j.scib.2020.05.015.

Demott, C. A., D. A. Randall, and M. Khairoutdinov, 2007: Convective precipitation variability as a tool for general circulation model analysis. *J. Climate*, **20**, 91–112, https://doi.org/10.1175/JCLI3991.1.

Dickinson, R. E., 1995: Land processes in climate models. *Remote Sens. Environ.*, **51**, 27–38, https://doi.org/10.1016/0034-4257(94)00062-R.

Ficklin, D. L., S. E. Null, J. T. Abatzoglou, K. A. Novick, and D. T. Myers, 2022: Hydrological intensification will increase the complexity of water resource management. *Earth's Future*, **10**, e2021EF002487, https://doi.org/10.1029/2021EF002487.

Fowler, H. J., and Coauthors, 2021: Anthropogenic intensification of short-duration rainfall extremes. *Nat. Rev. Earth Environ.*, **2**, 107–122, https://doi.org/10.1038/s43017-020-00128-6.

Frégier, Y., and J.-B. Gouray, 2021: Mind2Mind: Transfer learning for GANs. *Proc. 5th Int. Conf., GSI 2021*, Paris, France, Springer, 851–859, https://doi.org/10.1007/978-3-030-80209-7_91.

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

Gregory, D., 1997: Sensitivity of general circulation model performance to convective parametrization. *The Physics and Parameterization of Moist Atmospheric Convection*, R. K. Smith, Ed., Springer, 463–482.

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.

Ham, H., T. J. Jun, and D. Kim, 2020: Unbalanced GANs: Pretraining the generator of generative adversarial network using variational autoencoder. arXiv, 2002.02112v1, https://doi.org/10.48550/arXiv.2002.02112.

Hanesiak, J., and Coauthors, 2022: The severe multi-day October 2019 snow storm over southern Manitoba, Canada. *Atmos.–Ocean*, **60**, 65–87, https://doi.org/10.1080/07055900.2022.2060794.

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.

Hellinger, E., 1909: Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *J. Reine Angew. Math. (Crelles J.)*, **1909**, 210–271, https://doi.org/10.1515/crll.1909.136.210.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, https://doi.org/10.1002/qj.3803.

Hosking, J. R. M., 1990: L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *J. Roy. Stat. Soc.*, **52B**, 105–124, https://doi.org/10.1111/j.2517-6161.1990.tb01775.x.

Iotti, M., P. Davini, J. Von Hardenberg, and G. Zappa, 2025: RainScaleGAN: A conditional Generative Adversarial Network for rainfall downscaling. *Artif. Intell. Earth Syst.*, **4**, e240074, https://doi.org/10.1175/AIES-D-24-0074.1.

Jakob, C., and S. A. Klein, 2000: A parametrization of the effects of cloud and precipitation overlap for use in general-circulation models. *Quart. J. Roy. Meteor. Soc.*, **126**, 2525–2544, https://doi.org/10.1002/qj.49712656809.

Lanzante, J. R., K. W. Dixon, M. J. Nath, C. E. Whitlock, and D. Adams-Smith, 2018: Some pitfalls in statistical downscaling of future climate. *Bull. Amer. Meteor. Soc.*, **99**, 791–803, https://doi.org/10.1175/BAMS-D-17-0046.1.

Lauer, M., A. Rinke, I. Gorodetskaya, M. Sprenger, M. Mech, and S. Crewell, 2023: Influence of atmospheric rivers and associated weather systems on precipitation in the Arctic. *Atmos. Chem. Phys.*, **23**, 8705–8726, https://doi.org/10.5194/acp-23-8705-2023.

Lee, J., and S. Y. Park, 2025: WGAN-GP-based Conditional GAN (cGAN) with extreme critic for precipitation downscaling in a key agricultural region of the northeastern U.S. *IEEE Access*, **13**, 46 030–46 041, https://doi.org/10.1109/ACCESS.2025.3549443.

Lei, M., J. Li, L. Zhang, C. Deng, Y. Li, and J. He, 2022: Inconsistent frequency trends between hourly and daily precipitation during warm season in mainland of China. *Geophys. Res. Lett.*, **49**, e2022GL100277, https://doi.org/10.1029/2022GL100277.

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.

Lenderink, G., H. De Vries, E. Van Meijgaard, W. De Rooy, L. Van Ulft, V. Thompson, X. Qiu, and H. J. Fowler, 2025: A pseudo global warming based system to study how climate change affects high impact rainfall events. *Wea. Climate Extremes*, **49**, 100781, https://doi.org/10.1016/j.wace.2025.100781.

Li, Y., Z. Li, Z. Zhang, L. Chen, S. Kurkute, L. Scaff, and X. Pan, 2019: High-resolution regional climate modeling and projection over western Canada using a weather research forecasting model with a pseudo-global warming approach. *Hydrol. Earth Syst. Sci.*, **23**, 4635–4659, https://doi.org/10.5194/hess-23-4635-2019.

Liu, W., Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, 2017: A survey of deep neural network architectures and their applications. *Neurocomputing*, **234**, 11–26, https://doi.org/10.1016/j.neucom.2016.12.038.

Liu, Y., Z. Li, G. Cao, Q. Wang, Y. Li, and Z. Lu, 2025: MSRGAN: A multi-scale residual GAN for high-resolution precipitation downscaling. *Remote Sens.*, **17**, 2281, https://doi.org/10.3390/rs17132281.

Luo, N., Y. Guo, J. Chou, and Z. Gao, 2022: Added value of CMIP6 models over CMIP5 models in simulating the climatological precipitation extremes in China. *Int. J. Climatol.*, **42**, 1148–1164, https://doi.org/10.1002/joc.7294.

Maraun, D., 2013: Bias correction, quantile mapping, and downscaling: Revisiting the inflation issue. *J. Climate*, **26**, 2137–2143, https://doi.org/10.1175/JCLI-D-12-00821.1.

Milrad, S. M., J. R. Gyakum, E. H. Atallah, and J. F. Smith, 2011: A diagnostic examination of the eastern Ontario and western Quebec wintertime convection event of 28 January 2010. *Wea.*

Unauthenticated | Downloaded 10/09/26 09:56 AM UTC

Forecasting, **26**, 301–318, https://doi.org/10.1175/2010WAF2222432.1.

Mirza, M., and S. Osindero, 2014: Conditional Generative Adversarial Nets. arXiv, 1411.1784v1, https://doi.org/10.48550/arXiv.1411.1784.

O'Shea, K., and R. Nash, 2015: An introduction to Convolutional Neural Networks. arXiv, 1511.08458v2, https://doi.org/10.48550/arXiv.1511.08458.

Pan, B., K. Hsu, A. AghaKouchak, and S. Sorooshian, 2019: Improving precipitation estimation using convolutional neural network. *Water Resour. Res.*, **55**, 2301–2321, https://doi.org/10.1029/2018WR024090.

Papalexiou, S. M., and A. Montanari, 2019: Global and regional increase of precipitation extremes under global warming. *Water Resour. Res.*, **55**, 4901–4914, https://doi.org/10.1029/2018WR024067.

——, and A. Mamalakis, 2025: Machine unlearning: Bias correction in neural network downscaled storms. *J. Hydrol.*, **665**, 134689, https://doi.org/10.1016/j.jhydrol.2025.134689.

Petzka, H., A. Fischer, and D. Lukovnikov, 2018: On the regularization of Wasserstein GANs. *Int. Conf. on Learning Representations (ICLR)*, Vancouver, BC, Canada, OpenReview, https://www.researchgate.net/publication/320055586_On_the_regularization_of_Wasserstein_GANs.

Piani, C., G. P. Weedon, M. Best, S. M. Gomes, P. Viterbo, S. Hagemann, and J. O. Haerter, 2010: Statistical bias correction of global simulated daily precipitation and temperature for the application of hydrological models. *J. Hydrol.*, **395**, 199–215, https://doi.org/10.1016/j.jhydrol.2010.10.024.

Prein, A. F., and Coauthors, 2015: A review on regional convection-permitting climate modeling: Demonstrations, prospects, and challenges. *Rev. Geophys.*, **53**, 323–361, https://doi.org/10.1002/2014RG000475.

Radford, A., L. Metz, and S. Chintala, 2016: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv, 1511.06434v2, https://doi.org/10.48550/arXiv.1511.06434.

Rampal, N., P. B. Gibson, S. Sherwood, and G. Abramowitz, 2024: On the extrapolation of Generative Adversarial Networks for downscaling precipitation extremes in warmer climates. *Geophys. Res. Lett.*, **51**, e2024GL112492, https://doi.org/10.1029/2024GL112492.

Rasmussen, R., and Coauthors, 2011: High-resolution coupled climate runoff simulations of seasonal snowfall over Colorado: A process study of current and warmer climate. *J. Climate*, **24**, 3015–3048, https://doi.org/10.1175/2010JCLI3985.1.

Rasp, S., and N. Thuerey, 2021: Data-driven medium-range weather prediction with a resnet pretrained on climate simulations: A new model for WeatherBench. *J. Adv. Model. Earth Syst.*, **13**, e2020MS002405, https://doi.org/10.1029/2020MS002405.

Reddy, P. J., and R. Matear, 2025: Generating ensembles for convective-scale precipitation downscaling using generative machine learning. ESS Open Archive, https://doi.org/10.22541/essoar.174802957.72376154/v1.

——, ——, J. Taylor, M. Thatcher, and M. Grose, 2023: A precipitation downscaling method using a super-resolution deconvolution neural network with step orography. *Environ. Data Sci.*, **2**, e17, https://doi.org/10.1017/eds.2023.18.

Sankarasubramanian, A., and K. Srinivasan, 1999: Investigation and comparison of sampling properties of L-moments and conventional moments. *J. Hydrol.*, **218**, 13–34, https://doi.org/10.1016/S0022-1694(99)00018-9.

Saxena, D., and J. Cao, 2021: Generative Adversarial Networks (GANs): Challenges, solutions, and future directions. *ACM Comput. Surv.*, **54** (3), 1–42, https://doi.org/10.1145/3446374.

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang, and J. G. Powers, 2005: A description of the Advanced Research WRF version 2. NCAR Tech. Note NCAR/TN-468+STR, 100 pp., https://doi.org/10.5065/D6DZ069T.

Song, Y. H., M. S. Nashwan, E.-S. Chung, and S. Shahid, 2021: Advances in CMIP6 INM-CM5 over CMIP5 inm-CM4 for precipitation simulation in South Korea. *Atmos. Res.*, **247**, 105261, https://doi.org/10.1016/j.atmosres.2020.105261.

Tabari, H., S. M. Paz, D. Buekenhout, and P. Willems, 2021: Comparison of statistical downscaling methods for climate change impact analysis on precipitation-driven drought. *Hydrol. Earth Syst. Sci.*, **25**, 3493–3517, https://doi.org/10.5194/hess-25-3493-2021.

Tahara, R., Y. Hiraga, and S. Kazama, 2025: Climate change effects on the localized heavy rainfall event in northern Japan in 2022: Uncertainties in a pseudo-global warming approach. *Atmos. Res.*, **314**, 107780, https://doi.org/10.1016/j.atmosres.2024.107780.

Tefera, G. W., R. L. Ray, and A. M. Wootten, 2024: Evaluation of statistical downscaling techniques and projection of climate extremes in central Texas, USA. *Wea. Climate Extremes*, **43**, 100637, https://doi.org/10.1016/j.wace.2023.100637.

Thomassen, E. D., S. L. Thorndahl, C. B. Andersen, I. B. Gregersen, K. Arnbjerg-Nielsen, and H. J. D. Sørup, 2022: Comparing spatial metrics of extreme precipitation between data from rain gauges, weather radar and high-resolution climate model re-analyses. *J. Hydrol.*, **610**, 127915, https://doi.org/10.1016/j.jhydrol.2022.127915.

Tie, R., and Coauthors, 2024: Improve the downscaling accuracy of high-resolution precipitation field using classification mask. *Atmos. Res.*, **310**, 107607, https://doi.org/10.1016/j.atmosres.2024.107607.

Tirel, L., A. M. Ali, and H. A. Hashim, 2024: Novel hybrid integrated Pix2Pix and WGAN model with Gradient Penalty for binary images denoising. arXiv, 2407.11865v2, https://doi.org/10.48550/arXiv.2407.11865.

Trenberth, K. E., A. Dai, R. M. Rasmussen, and D. B. Parsons, 2003: The changing character of precipitation. *Bull. Amer. Meteor. Soc.*, **84**, 1205–1218, https://doi.org/10.1175/BAMS-84-9-1205.

Vandal, T., E. Kodra, and A. R. Ganguly, 2019: Intercomparison of machine learning methods for statistical downscaling: The case of daily and extreme precipitation. *Theor. Appl. Climatol.*, **137**, 557–570, https://doi.org/10.1007/s00704-018-2613-3.

Vosper, E., P. Watson, L. Harris, A. McRae, R. Santos-Rodriguez, L. Aitchison, and D. Mitchell, 2023: Deep learning for downscaling tropical cyclone rainfall to hazard-relevant spatial scales. *J. Geophys. Res. Atmos.*, **128**, e2022JD038163, https://doi.org/10.1029/2022JD038163.

Vrac, M., M. L. Stein, K. Hayhoe, and X.-Z. Liang, 2007: A general method for validating statistical downscaling methods under future climate change. *Geophys. Res. Lett.*, **34**, L18701, https://doi.org/10.1029/2007GL030295.

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.

Unauthenticated | Downloaded 10/09/26 09:56 AM UTC

Yang, J., S. He, and Q. Bao, 2021: Convective/large-scale rainfall partitions of tropical heavy precipitation in CMIP6 atmospheric models. *Adv. Atmos. Sci.*, **38**, 1020–1027, https://doi.org/10.1007/s00376-021-0238-4.

Zolina, O., C. Simmer, K. Belyaev, S. K. Gulev, and P. Koltermann, 2013: Changes in the duration of European wet and dry spells during the last 60 years. *J. Climate*, **26**, 2022–2047, https://doi.org/10.1175/JCLI-D-11-00498.1.

Zou, Y., Y. Wang, and X. Lu, 2023: Auto-encoding Generative Adversarial Networks towards mode collapse reduction and feature representation enhancement. *Entropy*, **25**, 1657, https://doi.org/10.3390/e25121657.

Unauthenticated | Downloaded 10/09/26 09:56 AM UTC