Earth's Future

RESEARCH ARTICLE

10.1029/2023EF003531

Key Points:

- Highest statistically downscaled spatio-temporal resolution multivariate ensemble currently available, consisting of 23 projection runs
- We downscaled precipitation, water vapor pressure, radiation, wind speed, and, maximum, mean and minimum temperature
- The methodology complied to a high degree with the three perfect prognosis assumptions and is scalable to other spatio-temporal resolutions

Correspondence to:

D. Quesada-Chacón,
dannell.quesada@tu-dresden.de;
dannell.quesada@pik-potsdam.de

Citation:

Quesada-Chacón, D., Baño-Medina, J., Barfus, K., & Bernhofer, C. (2023). Downscaling CORDEX through deep learning to daily 1 km multivariate ensemble in complex terrain. *Earth's Future*, *11*, e2023EF003531. https://doi.org/10.1029/2023EF003531

Received 25 JAN 2023
Accepted 30 JUN 2023

Author Contributions:

Conceptualization: Dánnell Quesada-Chacón, Jorge Baño-Medina, Klemens Barfus, Christian Bernhofer
Data curation: Dánnell Quesada-Chacón
Formal analysis: Dánnell Quesada-Chacón
Funding acquisition: Dánnell Quesada-Chacón, Klemens Barfus, Christian Bernhofer
Investigation: Dánnell Quesada-Chacón
Methodology: Dánnell Quesada-Chacón, Jorge Baño-Medina, Klemens Barfus
Project Administration: Dánnell Quesada-Chacón
Resources: Dánnell Quesada-Chacón
Software: Dánnell Quesada-Chacón

Downscaling CORDEX Through Deep Learning to Daily 1 km Multivariate Ensemble in Complex Terrain

Dánnell Quesada-Chacón[1], Jorge Baño-Medina[2], Klemens Barfus[1], and Christian Bernhofer[1]

[1]Institute of Hydrology and Meteorology, TU Dresden, Dresden, Germany, [2]Instituto de Física de Cantabria (IFCA), CSIC–Universidad de Cantabria, Santander, Spain

Abstract High spatio-temporal resolution near-surface projected data is vital for climate change impact studies and adaptation. We derived the highest statistically downscaled resolution multivariate ensemble currently available: daily 1 km until the end of the century. Deep learning models were employed to develop transfer functions for precipitation, water vapor pressure, radiation, wind speed, and, maximum, mean and minimum temperature. Perfect prognosis is the particular statistical downscaling methodology applied, using a subset of the ReKIS data set for Saxony as predictands, the ERA5 reanalysis as during-training predictors and the CORDEX-EUR11 ensemble as projected predictors. The performance of the transfer functions was validated with the VALUE framework, yielding highly satisfactory results. Particular attention was given to the three major perfect prognosis assumptions, for which several tests were carried out and thoroughly discussed. From the latter, we corroborated their fulfillment to a high degree, thus, the derived projections are considered adequate and relevant for impact modelers. In total, 18 runs for RCP85, 1 for RCP45, and 4 for RCP26 were downscaled under both stochastic and deterministic approaches. This multivariate ensemble could drive more accurate and diverse impact studies in the region. Generally, the projected climatologies are in agreement with coarser resolution projections. Nevertheless, statistical particularities were observed for some projections, thus, a list of caveats for potential users is given. Due to the scalability of the presented methodology, further possible applications with additional datasets are proposed. Lastly, several potential improvement prospects are discussed toward the ideal subsequent iteration of the perfect prognosis statistical downscaling methodology.

Plain Language Summary There is a great worldwide demand for high spatio-temporal resolution projections to develop climate change adaptation and mitigation schemes. Despite recent improvements, the resolution of both global and regional climate models is still too coarse to properly represent local variability, particularly in complex terrains. Depending on the application, impact modelers and decision makers require kilometer-scale projections, with a minimum daily temporal resolution, of near-surface variables. To fill this information gap, we employed artificial intelligence algorithms to downscale, to a novel daily 1 km resolution, a projection ensemble until the end of the century consisting of precipitation, water vapor pressure, radiation, wind speed, and, maximum, mean and minimum temperature. The ensemble comprises 18 runs of the business-as-usual worst-case scenario (RCP85), 1 run of the stabilization scenario (RCP45), and 4 of the optimistic low-emissions scenario (RCP26). The main assumptions of the methodology were thoroughly tested and discussed. The validation carried out yielded highly satisfactory results. Thus, we consider the projections to be adequate and relevant for impact studies. The region studied is located in Saxony (Germany), still, the methodology shown is potentially applicable anywhere in the world.

1. Introduction

According to IPCC (2021), the global average surface temperature for 2011–2020 has increased by 1.09°C compared to 1850–1900 and is expected to reach values of ∼3–5°C by the end of the century in the worst case scenario, as described by the Representative Concentration Pathway (RCP) 8.5, from the Coupled Model Intercomparison Project Phase 5 (CMIP5, Taylor et al., 2012). As a consequence, a large number of different socio-economic sectors and political institutions require climate information to elaborate adaptation and mitigation plans for climate change at global, regional and local scales.

At global scales, General Circulation Models (GCMs) are the main tools used to project the effect of various forcing scenarios on the climate system, for example, different greenhouse gas concentration trajectories. GCMs are numerical models that represent the physical processes in the atmosphere, land, and ocean.

Supervision: Jorge Baño-Medina, Klemens Barfus, Christian Bernhofer
Validation: Dánnell Quesada-Chacón
Visualization: Dánnell Quesada-Chacón
Writing – original draft: Dánnell Quesada-Chacón
Writing – review & editing: Dánnell Quesada-Chacón, Jorge Baño-Medina, Klemens Barfus, Christian Bernhofer

Regardless of the recent developments on GCMs, their coarse spatial resolution (∼100–200 km) and large regional biases render their output unsuitable for regional or local climate change impact studies (Maraun & Widmann, 2018). Downscaling techniques, dynamical and statistical, are used to improve the resolution of the GCM output. Regional Climate Models (RCMs) are employed to dynamically downscale GCM output by using the latter as boundary conditions to drive higher-resolution nested models. The Coordinated Regional Climate Downscaling Experiment (CORDEX, 2021) provides RCM output with multiple spatio-temporal resolutions. Still, the highest resolution available (0.11° only for Europe) is unable to adequately represent near-surface variables, especially for topographically complex areas. Thus, the application and location define the spatio-temporal resolution required, which is often limited by the coarser data available. Moreover, Katragkou et al. (2015) demonstrated that such variables exhibit significant systematic biases within EURO-CORDEX models (conducted with a spatial resolution of 0.44°) as a result of the employed parametrization schemes. There are many efforts underway to generate kilometer-scale (∼1–2 km grid) global climate models (Schär et al., 2020), due to the performance improvements in the impact models that such higher resolutions convey in comparison to coarser ones (Quintero et al., 2022). The latter would require explicitly resolving small-scale convective cloud processes, enormous computing power, and, important efforts to adapt the existing GCMs and RCMs code to the newest, GPU-based, supercomputer architectures (Schär et al., 2020).

Statistical downscaling methods represent a cost-effective approach to build high-resolution datasets by establishing *transfer functions* between large-scale variables (predictors) and regional- or local-scale variables (predictands), as described in Maraun and Widmann (2018). There are two major statistical downscaling schemes, that is, perfect prognosis (PP) and model output statistics. PP models are calibrated with predictands (observations) and predictors (taken from reanalysis data, generally atmospheric variables) that hold a strict temporal correspondence. There are three major assumptions related to the PP approach, which can be summarized in: (a) the predictors need to be realistically and bias-free simulated, (b) the predictors should explain a large portion of the variability, and (c) the influence of the predictors on the predictands needs to be sensibly modeled, allowing at least moderate extrapolations for non-observed climate (Maraun & Widmann, 2018). The combination of (b) and (c) is also known as the time-invariance assumption. On the other hand, model output statistics does not require temporal correspondence and defines the transfer function between model and observed data (generally the same variable) to post-process model data, which bias-corrects it.

Recently, statistical downscaling has seen major improvements with a growing number of applications. Still, most of the studies statistically downscale precipitation and/or temperature from GCM output to station data (Gutiérrez et al., 2019; Olmo et al., 2022) or to another grid, which is generally rather coarse to be employed in local-scale impact models (Vandal et al., 2018; Baño-Medina et al., 2022, ∼12.5 km and 0.5°, respectively). Moreover, several global high-resolution datasets have been created with the CHELSA mechanistic statistical downscaling algorithm (Karger et al., 2017), such as Karger et al. (2020), with ∼5 km monthly projections of precipitation and temperature. Additionally, very few studies emphasize on other near-surface variables needed to characterize important impacts of climate change at the regional scale, for example: (a) humidity (e.g., Huth, 2005; Pierce & Cayan, 2016), needed for crop impact models, (b) radiation (Rivington et al., 2008), meaningful for a variety of impact models and decarbonization through projections of solar energy production, (c) wind speed (e.g., Höhlein et al., 2020; Ramon et al., 2021), which is relevant, among others, for wind power energy production, and (d) air pollutants such as particulate matter (Wise, 2009) and ground-level ozone (Hertig et al., 2023; Wise, 2009), which are key for ecosystems and human health.

Moreover, projections of variables not limited to precipitation and/or temperature are offered by only a very limited amount of studies, for example, Lange (2019, 0.5° daily, 10 variables) and Brun et al. (2022, ∼1 km projected 30 years averages of bioclimatic and agriculture-relevant variables), both statistically downscaled with a model output statistics approach. Projections of further variables are needed for more diverse and specialized impact models. Lately, novel methods for multivariate bias correction have been developed and tested, for example, probability density functions transformations and bivariate copulas (Cannon, 2018; François et al., 2020, 2021; Vrac, 2018). Nevertheless, these studies are focused on precipitation and temperature only. Spatio-temporal coherent multivariate projections are necessary to assess the risk of future compound events, where the combined effect of different drivers across several spatial and temporal scales can cause major impacts (Zscheischler et al., 2018).

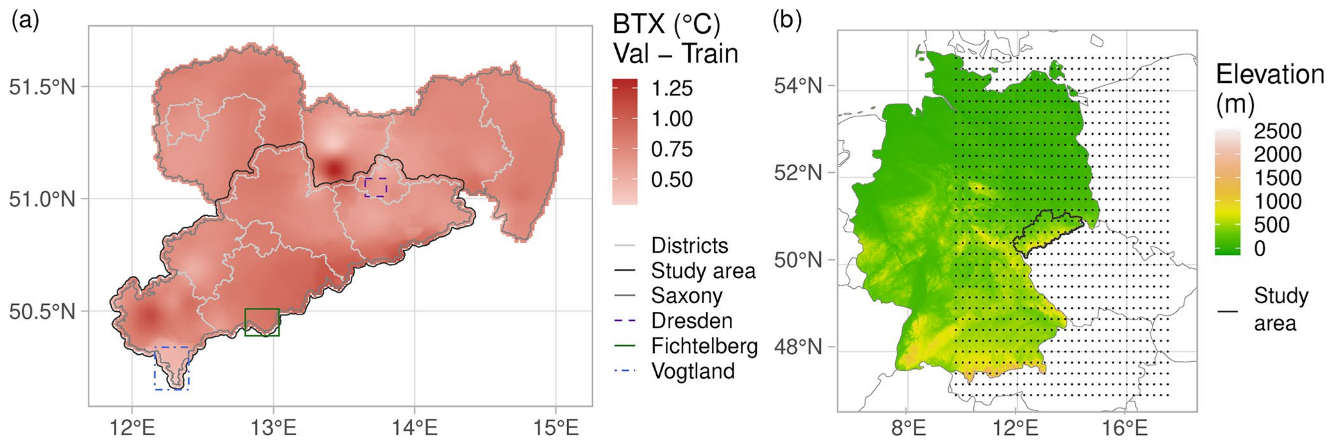

Figure 1. Location of the study area and the predictor domain. (a) The bias of maximum temperature (BTX) between training and validation periods for the whole *ReKIS* domain for Saxony. The study area is inside the darker gray line. The subregions of Dresden, Fichtelberg and Vogtland will be used for further analysis. (b) Topography of Germany, the center of the ERA5 sub-domain pixels (marked by dots, 32 by 32) used for the predictors and the study area.

Due to the enormous computational requirements of RCMs to simulate climate at a convection-permitting scale and the limited informative power of very coarse variables to establish empirical relationships directly with the local scale, some studies have proposed a hybrid dynamical-statistical approach which takes advantage of both downscaling families (e.g., Li et al., 2020; Quesada-Chacón et al., 2020). This hybrid approach employs higher resolution RCM output, which shows generally lower biases than GCMs when compared to observations (Sørland et al., 2018), to further statistically downscale it to the local scale. In the case of the two hybrid dynamical-statistical studies, station data was downscaled. However, statistical downscaling of projected daily RCM output to higher-resolution gridded data remains, to our knowledge, unpublished, and could prove to be highly beneficial for impact models.

The aim of this paper is to downscale RCM output to a daily 1 km multivariate (precipitation, water vapor pressure, radiation, wind speed, and, maximum, mean and minimum temperature) projection ensemble until the end of the century employing the PP methodology through deep learning (DL). Quesada-Chacón et al. (2022) proved that DL is capable to learn complex atmospheric patterns for downscaling tasks in the region of interest. There, the theoretical, methodological, and computational bases were established and tested only for precipitation during past conditions, bearing scalability in mind, that is, also suitable for different: temporal and spatial resolutions, and regions. This paper builds upon the methods and containerized software environment previously developed in Quesada-Chacón et al. (2022), while introducing several novel aspects. Specifically, our work encompasses (a) extending the workflow to include additional near-surface variables, (b) expanding the methods into the projection domain, and (c) assessing the quality of the produced data set by scrutinizing the fulfillment of the three major PP assumptions and conducting a comparative analysis to evaluate the climate change signal in relation to coarser resolutions. The Regional Climate Information System for Saxony, Saxony-Anhalt, and Thuringia data set (ReKIS, 2021) is used as predictands, the *ERA5* reanalysis (Hersbach et al., 2020) as training predictor, and, the CORDEX EUR11 (Jacob et al., 2014) as the projected predictors. We hope to provide the information needed to drive more accurate and diverse impact studies for the study region through this multivariate ensemble.

The rest of this article is organized as follows: Section 2 displays the study region, the datasets employed as predictands and predictors, and, the methodological approach. Section 3 presents the results, as well as the discussion. Lastly, the summary of the present work and outlook for future research are shown in Section 4.

2. Data and Methods

2.1. Study Area

The present study region (see Figure 1a) includes the Ore Mountains/Vogtland Nature Park (the longest nature park in Germany), the Saxon Switzerland National Park and a large portion of the flatlands toward the north

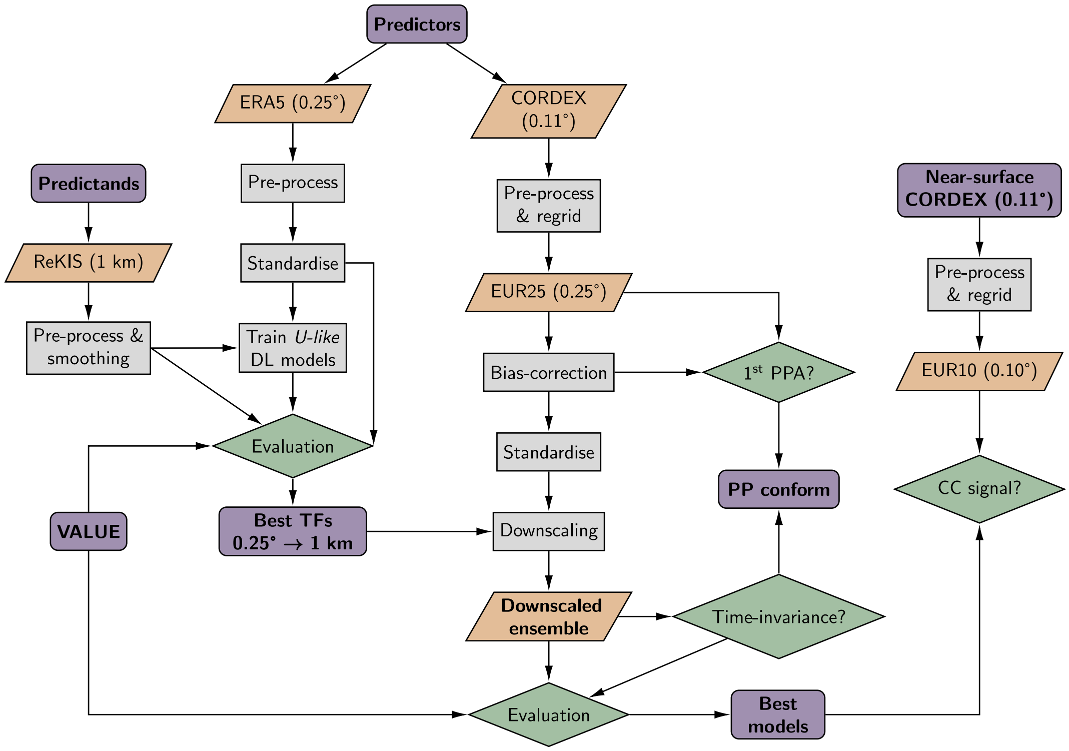

Figure 2. Methodological approach employed. PPA, PP assumption; TF, transfer function; CC, climate change.

of Saxony, including its capital, Dresden. This region is an extension of the one employed in Quesada-Chacón et al. (2022) since we sought to test the scalability of the methodology and to cover a larger and more relevant region for impact modelers. The subregions shown in Figure 1a were selected as representative, that is, Dresden exemplifies the climate of the northern flatlands, Fichtelberg (elevation 1,215 m) depicts the climate of the highest elevations with complex topography, while Vogtland exhibits intermediate climatic conditions.

2.2. Predictands

A subset of the ReKIS (2021) gridded data set for the Free State of Saxony is used as predictand, with a spatio-temporal resolution of daily 1 km. This data set uses station data from the Czech Hydrometeorological Institute (CHMI) and the German Meteorological Service (Deutscher Wetterdienst [DWD]) as sources to produce gridded data through geostatistic interpolation such as thin plate splines (Körner et al., 2022). Seven variables were taken into account for downscaling in the present paper, which are: minimum (TN [°C]), mean (TM [°C]) and maximum temperature (TX [°C]), precipitation (Pr [mm · day^{-1}]), radiation (Rn [kW · h · m^{-2}]), wind speed (WS [m · s^{-1}]) and relative humidity. The latter was converted to water vapor pressure (Pw [hPa]) according to Huang (2018) to avoid artifacts in situations where the downscaled values could be above 100%. The aforementioned abbreviations and units for the variables will be employed hereafter. The predictands were cropped to the focus region satisfying the information needs of multiple impact models. Additionally, since the variables in the original ReKIS data set were rounded to one decimal place, random noise of a lesser order of magnitude was added as a measure to *de-discretize* the observations of continuous nature to derive the final predictands. The latter will in turn ease the training process of the DL models. The last two steps are illustrated by the "Pre-process and smoothing" box of Figure 2. Note the two arrows coming out of this box, portraying the *training* subset (1979–2005) and the *validation* subset (2006–2015).

Table 1
Details of the Selected EUR11 Model Output

		RCM				
			Member			
Experiment	GCM	Version	CLMcom-ETH-COSMO-crCLIM-v1-1	ICTPRegCM4-6	CNRM-ALADIN63	Total
Historical	CNRM-CERFACS-CNRM-CM5	1	1			18
		2		1	1	
	ICHEC-EC-EARTH	1	1, 3, 12	12		
	MOHC-HadGEM2-ES	1	1	1	1	
	MPI-M-MPI-ESM-LR	1	1, 2, 3	1	1	
	NCC-NorESM1-M	1	1	1	1	
RCP26	CNRM-CERFACS-CNRM-CM5	2			1	4
	MOHC-HadGEM2-ES	1		1		
	MPI-M-MPI-ESM-LR	1		1		
	NCC-NorESM1-M	1		1		
RCP45	CNRM-CERFACS-CNRM-CM5	2			1	1
RCP85	CNRM-CERFACS-CNRM-CM5	1	1			18
		2		1	1	
	ICHEC-EC-EARTH	1	1, 3, 12	12		
	MOHC-HadGEM2-ES	1	1	1	1	
	MPI-M-MPI-ESM-LR	1	1, 2, 3	1	1	
	NCC-NorESM1-M	1	1	1	1	

2.3. Predictors

Two datasets were employed as predictors in the present paper. First, a subset of the ERA5 reanalysis (Hersbach et al., 2020), from 1979 to 2005, cropped to a 32 by 32 pixels domain (see Figure 1b), was used to train the models under perfect-prognosis conditions (Maraun & Widmann, 2018). The subset of ERA5, from 2006 to 2015 was used to validate the models. Analogously, note the two arrows coming out of the "Standardize" box in Figure 2. Then, the EURO-CORDEX data set (Jacob et al., 2014) is coupled with the trained models to obtain the multivariate projected ensemble. For this procedure to properly work, the predictor sets for both datasets need to contain the same variables and the same resolution.

Therefore, a metadata screening of the whole EURO-CORDEX data set was initially conducted to determine which variables to employ, trying to maximize the ensemble size without compromising key predictors, since not all GCM-RCM combinations (GRCMCs) offer the same variables. Originally, it was intended to use the CORDEX EUR22 data set, which has a spatial resolution of 0.22°, closer to the one of ERA5 (0.25°). Nevertheless, the aforementioned screening yielded that there was limited model output available for EUR22. Alternatively, it was found that the EUR11 data set, with a native spatial resolution of 0.11°, contained a substantial number of complying ensemble members. Based on similar studies (Baño-Medina et al., 2020, 2022; Quesada-Chacón et al., 2022), the selected variables are: zonal and meridional wind (u and v, respectively), temperature (t), geopotential (z), and specific humidity (q) at the 925, 850, 700, 500 and 200 hPa levels, and total cloud fraction (tcc), for a total of 26 predictors.

The selected EUR11 models, along with the details of their runs, are presented in Table 1. Thus, the number of ensemble members experiment-wise is: 18 for the historical period, four for RCP26, one for RCP45 and 18 for RCP85, with a combined size of more than 17 TB. The raw EUR11 data was downloaded, upscaled (using bilinear interpolation) and processed to match the grid and the units of the ERA5 subdomain. The missing values found in the data set were filled with the nearest neighbor, employing the `setmisstonn` operator of Climate Data Operators (`CDO`, Schulzweida, 2021). The upscaled and filled predictor ensemble is hereafter referred to as *EUR25*, as pictured in Figure 2. Consequently, all the 41 different GRCMCs will be coupled with the trained models to obtain projections for all the seven predictands.

Moreover, to comply with the PP assumption that the predictors are realistically and bias-free simulated in present climate (Maraun & Widmann, 2018), we employ a bias-adjustment technique on the predictors to enhance the distributional similarity between the GCM and reanalysis fields. Specifically, we adopt the Scaling Delta Mapping (SDM) technique following the approach outlined in Baño Medina et al. (2022). This technique preserves the monthly delta change of the predictors (i.e., the climatological difference between the future and historical periods for a given month), while replacing the simulated seasonal cycle with the reanalysis data for a reference period. For instance, for January 1992, we subtract the RCM's January monthly mean from the reference period of the RCM simulation and then add the equivalent mean from the reanalysis data set for the same reference period. The aforementioned PP assumption is then tested ("first PPA?" in Figure 2) for both bias-corrected and *raw* EUR25.

2.4. Transfer Functions

Based on the characteristics of the best performing statistical downscaling models or transfer functions of Quesada-Chacón et al. (2022), various DL architectures were tested for each predictand ("Train U-like DL models" in Figure 2). Both *U-Net* (Ronneberger et al., 2015) and *U-Net++* (Zhou et al., 2018) were tested with three and four layers and 64 and 128 filters on the first layer. Thus, eight different architectures were trained per predictand. Other *hyperparameters* previously tested were fixed to: one filter on the last *ConvUnit*, *dropout* of 0.25, *Leaky ReLu* with $\alpha = 0.3$ as activation function inside the *U* structures and the last *ConvUnit*, *batch normalization* both inside the *U-like* models and the last *ConvUnit*, *batch size* of 512, *Adam* as optimizer with a *learning rate* of $5 \cdot 10^{-4}$, *patience* of 125 epochs, a *validation split* of 0.1, and, 7777 as *random seed number*. As in Quesada-Chacón et al. (2022), the models were trained within a containerized environment (Quesada-Chacón, 2023d, v2.0.0) on a single NVIDIA A-100 GPU from the *Alpha Centauri* sub-cluster of the Center for Information Services and High Performance Computing (ZIH) of the Technische Universität Dresden.

Since new predictands are to be downscaled, whose cumulative distribution functions vary significantly, different *loss functions* were tested per predictand. Particularly, the negative log-likelihood of several probability distribution functions (PDFs) were the functions to optimize, besides the root-mean-squared-error (RMSE), which was added for comparison. A clear advantage of fitting PDFs to the predictands is the possibility of obtaining both deterministic (expected value or mean) and stochastic values. The latter type is desirable to analyze extreme values, which is necessary under climate change conditions (Maraun & Widmann, 2018). The PDFs tested are: Bernoulli Gamma (BG) and Gaussian for Pr; Gamma and Gaussian for Pw, Rn and WS; and Gaussian for TM, TN, and TX. Thus, several combinations of PDFs and architectures were trained individually per predictand. As a remark, additional loss functions were tested during the iteration process toward the present paper, that is, Bernoulli Log-normal for Pr, Log-normal for Pw, Rn and WS, and Weibull for WS; still, these loss functions did not perform as well as the above-mentioned ones.

Furthermore, considering the risk of future compound events (Zscheischler et al., 2018), another set of models was trained for all the predictands simultaneously, that is, seven branches, instead of one, derive at the end of the *U-like* structure, each with one filter on the last *ConvUnit*. The rationale behind this experiment is an attempt to maintain more strictly the daily relationship between the spatio-temporal features deducted from the atmospheric predictors by the *U-like* architectures, and the *coherence* among the predictands. Thus, each of the eight different architectures was trained for the different aforementioned loss functions, and also, trained independently with each predictand as well as with the full set of them.

2.5. Validation and Evaluation

The different combinations of transfer functions are then evaluated employing a subset of the VALUE framework metrics (Gutiérrez et al., 2019; Maraun et al., 2014, see Figure 2). For this, a validation data set from ERA5 is employed, which spans from 2006 to 2015. This subset was not used during training, therefore is completely independent. Then, the metrics of the individual models per metric per predictand were calculated and ranked, from which an overall ranking is derived per predictand from which the best-performing transfer functions are selected (as illustrated in Figure 2). The metrics are shown in Table 2 and were selected bearing in mind relevant aspects of the predictands, such as extremes, temporal characteristics and spatio-temporal coherence.

Table 2
Subset of the VALUE Metrics Employed to Validate the Performance of the Transfer Functions for the Different Predictands

Metric	Pr	Pw	Rn	TM`	TN	TX	WS	Description
BAC1		D	D	D	D	D	D	Lag-1 autocorrelation
BColdAMS					D			Median of the annual cold (<10th percentile) spell maxima
BDryAMS	D							Median of the annual dry (<1 mm) spell maxima
BFA20					D			Relative frequency of days >20°C (Tropical nights)
BFA25						D		Relative frequency of days >25°C (Summer days)
BFB0					D	D		Relative frequency of days <0°C (Ice days for TX; Frost days for TN)
BM	D	D	D	D	D	D	D	Mean
BP02	D	D, S	D, S	D, S	D, S	D, S	D	2nd percentile
BP98	D, S	D, S	D, S	D, S	D, S	D, S	D, S	98th percentile
BSDII	D							Mean wet-day (≥1 mm) precipitation (Simple Day Intensity Index)
BWarmAMS						D		Median of the annual warm (>90th percentile) spell maxima
BWetAMS	D							Median of the annual wet (≥1 mm) spell maxima
KSS	D	D	D	D	D	D	D	Kolmogorov–Smirnov statistic
Pearson		D	D	D	D	D		Pearson correlation
RSD	D	D	D	D	D	D	D	Ratio of the standard deviations
RMSE	D	D	D	D	D	D	D	Root Mean Square Error
Spearman	D						D	Spearman correlation

Note. (a) In *Metric*, the first letter *B* stands for *bias*. (b) The type of run used for the calculation is given below the predictands, that is, *D* stands for *deterministic* and *S* for *stochastic*.

Subsequently, the bias-corrected EUR25 ensemble of predictors is coupled with the best-performing transfer functions per predictand to obtain the daily *downscaled ensemble* (see Figure 2) until the end of the century. All the downscaled values were calculated under both deterministic and stochastic conditions. The historical period serves as pseudo-observations, as in similar downscaling approaches (Baño-Medina et al., 2022; Quesada-Chacón et al., 2020; San-Martín et al., 2017), from which the performance of the different GRCMCs can be further analyzed. We employed a subset of the metrics shown in Table 2 to generate a ranking for the GRCMCs, to highlight the best-performing ones for potential users. Moreover, to assess the extrapolation skill of the models (time-invariance assumption) and the quality of the projections, two future subperiods were established to examine the climatologies of the highest ranked GRCMs, that is, near future (NF, 2021–2050) and far future (FF, 2071–2100). The corresponding near-surface variables from EUR11 were downloaded and processed to a 0.1° regular grid (since each RCM has a different non-regular grid), hereafter referred to as EUR10, to compare the climate change signals of the downscaled projections. The above-mentioned steps are illustrated by Figure 2 as well.

3. Results and Discussion

3.1. Transfer Functions Performance

DL models were trained for the different combinations of architecture, loss function and predictand. The VALUE framework was used to validate and quantify the performance of the models, which allows us to rank them and objectively select the best-performing ones. Figure 3 shows the common subset among the predictands of the validation metrics presented in Table 2 for the top-ranked models per predictand. For coherence, the nomenclature used for the architectures follows the one used in Quesada-Chacón et al. (2022), that is, type of *U-like* structure (*U* for *U-Net* and *Upp* for *U-Net++*), number of layers inside the *U* structures (3 and 4), number of starting filters (64 and 128) for the *U* structures, number of filters of the last *ConvUnit*, and a *boolean* for *batch normalization* in the last *ConvUnit*, in that order. Due to the experiments conducted in the previous study, the latter two hyperparameters were set constant to 1 and `TRUE`, respectively. Note that all the metrics shown in Table 2 were calculated and used for ranking the models.

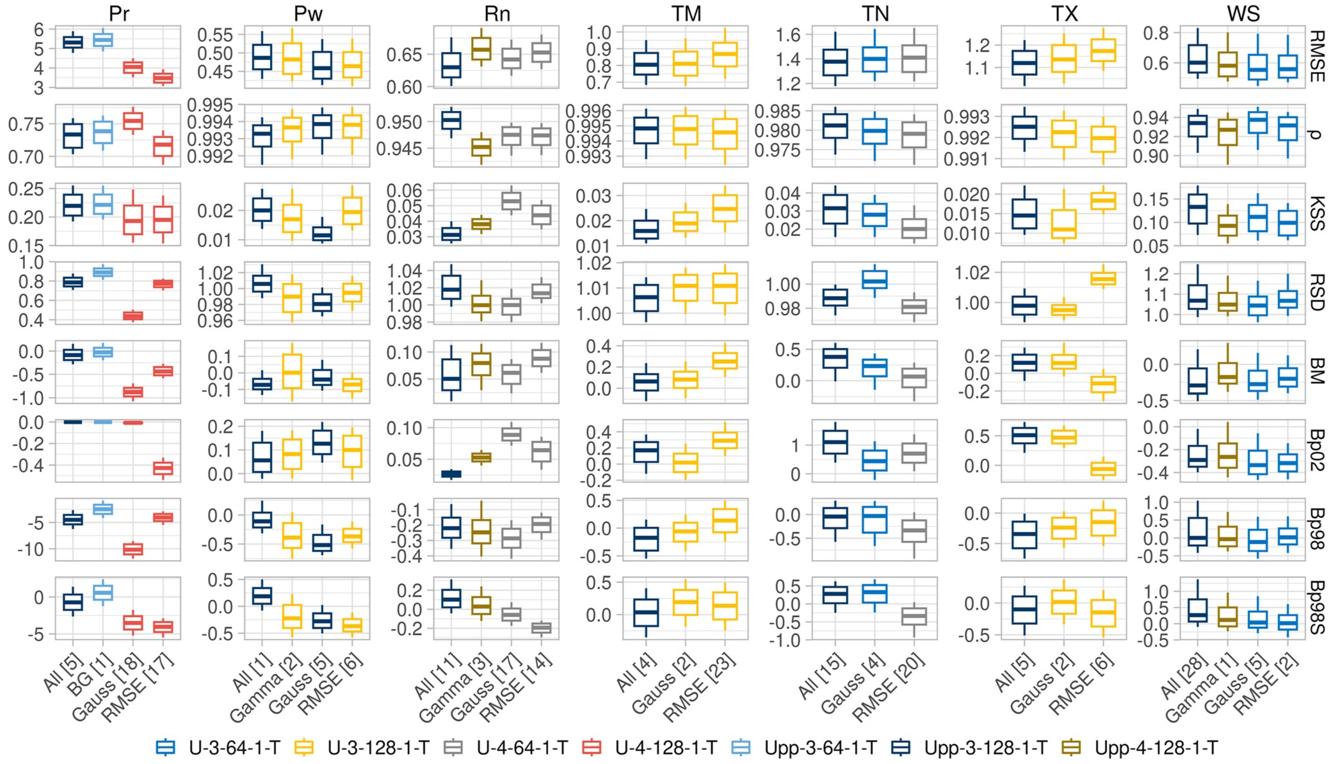

Figure 3. Box plot array of validation metrics for the best-performing methods per training loss scenario. The scenario where the same transfer function was trained multivariately is labeled as *All*. The predictands are ordered column-wise and the metrics row-wise. The abbreviations of the metrics are taken from Table 2, except for ρ, which depending on the predictand is either Pearson or Spearman correlation. Except for Bp98S, all the metrics shown were calculated from the deterministic results. The boxes comprise the 25th and 75th percentile and the median, and the whiskers the 10th and 90th percentile. The numbers inside the brackets show the overall ranking per predictand.

In general, the best-ranking individual performing loss functions are: BG for Pr, Gamma for Pw, Rn and WS, and Gauss for TM, TN and TX. As expected, the most challenging predictand to accurately model is Pr, still, the performance of the best-ranked model, *Upp-3-128-1-T* with BG (ranked #1), is highly satisfying. It is noticeable that for RMSE, the models trained with either Gauss or with RMSE performed better than BG. This outcome is expected, particularly for RMSE since that is precisely the metric optimized during training. This case has the disadvantage that no stochastic values can be obtained from it, for which its deficiencies can more clearly be noticed when compared with the stochastic values for the extreme values (Bp98S) from BG, and also has flaws in characteristics such as the dry and wet spells (not shown). Note that BM for Pr with the validation data set is fairly close to zero.

The other predictands had in most of the shown metrics in Figure 3 performances close to the ideal ones, that is, one for ρ and RSD, and zero for the rest. Yet, it is noticeable that no dramatic performance changes, as in Pr, occurred among the different loss functions for the shown metrics. The latter can also be understood by the smaller differences among the rankings of the best-performing combination of architectures and loss functions per individual predictand, for example, BG (ranked #1) and Gauss (#17) for Pr, Gamma (#1 and #1) and Gauss (#3 and #5) for Pw and WS, respectively.

In the case of the multivariately trained transfer function, the one shown in Figure 3 corresponds to the one that had the best joint performance, which has an architecture of *Upp-3-128-1-T*. The individual loss functions in this case are the previously mentioned best-ranking ones individually. Despite having a reasonable performance in most of the predictands, the ranks of TN (#21) and WS (#27) indicate their lesser performance in some of the metrics, for example, BM and Bp02 for TN, and, KSS and Bp98S for WS. It was noted that some multivariately trained models became rather specialized in some of the predictands (Pr, Pw, Rn, TM and TX ranking ≤#5), while performing poorly in the others, which is undesirable.

The rationale behind the multivariately trained transfer functions scenario is to preserve the relationship among the predictands and corresponding atmospheric conditions with a particular awareness of extreme compound events, which is certainly worth inspecting. Nevertheless, this scenario will not be sought hereafter since: (a) the individual models achieved considerably higher performances, (b) despite being trained simultaneously, the loss functions are individually calculated, and (c) the pursued *coherence* could be lost for univariately calculated stochastic results, as hereby implemented. The latter could even increase the biases of multivariate hazard estimates (Zscheischler et al., 2019). Posterior efforts for a similar scenario suggest a multivariate copula approach (François et al., 2020) which would allow, ideally, interdependent stochastical results, thus addressing (b) and (c). A comprehensive validation framework, comparable to VALUE, would also be needed for multivariate characteristics. Several efforts are underway toward the latter (e.g., Bevacqua et al., 2021; Zscheischler et al., 2020), still, these procedures are quite new and do not offer yet a thorough set of multivariate indexes as for example, VALUE. For the time being, such further investigations are beyond the scope of the present paper but could improve the analysis of future extreme compound events under climate change.

Despite the aforementioned limitations of the methodology, we find the performance of the best-ranked individual transfer functions suitable to generate the downscaled data set from the EUR25 bias-corrected ensemble.

3.2. Bias-Correction of Predictors

To evaluate and show the performance of the bias-correction method, we produced portrait plots (Gleckler et al., 2008) of the relative mean bias, the ratio of the variances, and, Kolmogorov–Smirnov statistic and *p*-values, for both bias-corrected and raw EUR25 data (see Figure 4). The perfect score for the metrics shown is zero (included in the white triangles), and, no symbol in Figure 4c is the ideal scenario. This figure aims to provide measures to the first assumption: "perfect prognosis means that the predictors have to be realistically and bias-free simulated in present climate" (Maraun & Widmann, 2018). Furthermore, an abbreviated nomenclature is used for the models (based on Table 1) in the following figures, where the space is limited, that is, *GCM_RCM_RCM-version_member*. Abbreviations: CNRM-CERFACS-CNRM-CM5 → CM5, ICHEC-EC-EARTH → EC, MOHC-HadGEM2-ES → ES, MPI-M-MPI-ESM-LR → MPI, NCC-NorESM1-M → Nor, CNRM-ALADIN63 → ALA, CLMcom-ETH-COSMO-crCLIM-v1-1 → COS, and ICTP-RegCM4-6 → Reg.

Generally, raw (lower right triangles) meridional wind is the atmospheric variable with the highest biases when compared to ERA5. Other noteworthy cases are the characteristics of geopotential and temperature for the raw data, where almost no bias is found but rather high differences in their variances are observed. Remarkably, most of the bias-corrected permutations of variables and GRCMCs (upper left triangles) show metric values very close to the ideal ones (white colored triangles) for both (a) and (b), with exceptions in several q200 combinations but still quite low differences, that is, from 1% to 10% for (a), and from 1.01 to 1.10 for (b). Hence, we can sustain to a great degree the bias-free portion of the first assumption for the predictor sets.

Figure 4c shows a lower amount of white triangles, even for bias-corrected variables. Since the bias-correction method is based on mean and variance, it is expected to have better performance in those aspects. Moreover, the deficiencies in the Kolmogorov–Smirnov statistic are linked to the limitations of the bias-correction method employed. The observed differences for some predictors, for example, several cases of tcc have between 5% and 11% for KSS, could cause systematic biases in the downscaled values, particularly in the predictands for which tcc is highly relevant, like Pr and Rn. This bias in tcc has also been detected in other studies (e.g., Katragkou et al., 2015).

Furthermore, the symbols in Figure 4c provide a measure of the number of pixels in which the null hypothesis of the Kolmogorov–Smirnov test can be rejected with a *p*-value < 0.05. Thus, no symbol over a white upper left triangle would be the ideal case per variable–GRCMC, which could be a criterion for the "realistic" bit of the aforementioned PP assumption. The latter coupled with values very close to the ideal one (white triangles) in the other two metrics could be interpreted as fulfilling it. Nevertheless, not a single GRCMC (column-wise) show white upper triangles for all the variables in all three metrics, which should be pondered for the results. Furthermore, many triangles with medium values of KSS show no symbol above them, like tcc, which could be interpreted as false positives. Quantile mapping approaches could be used to improve this particular aspect

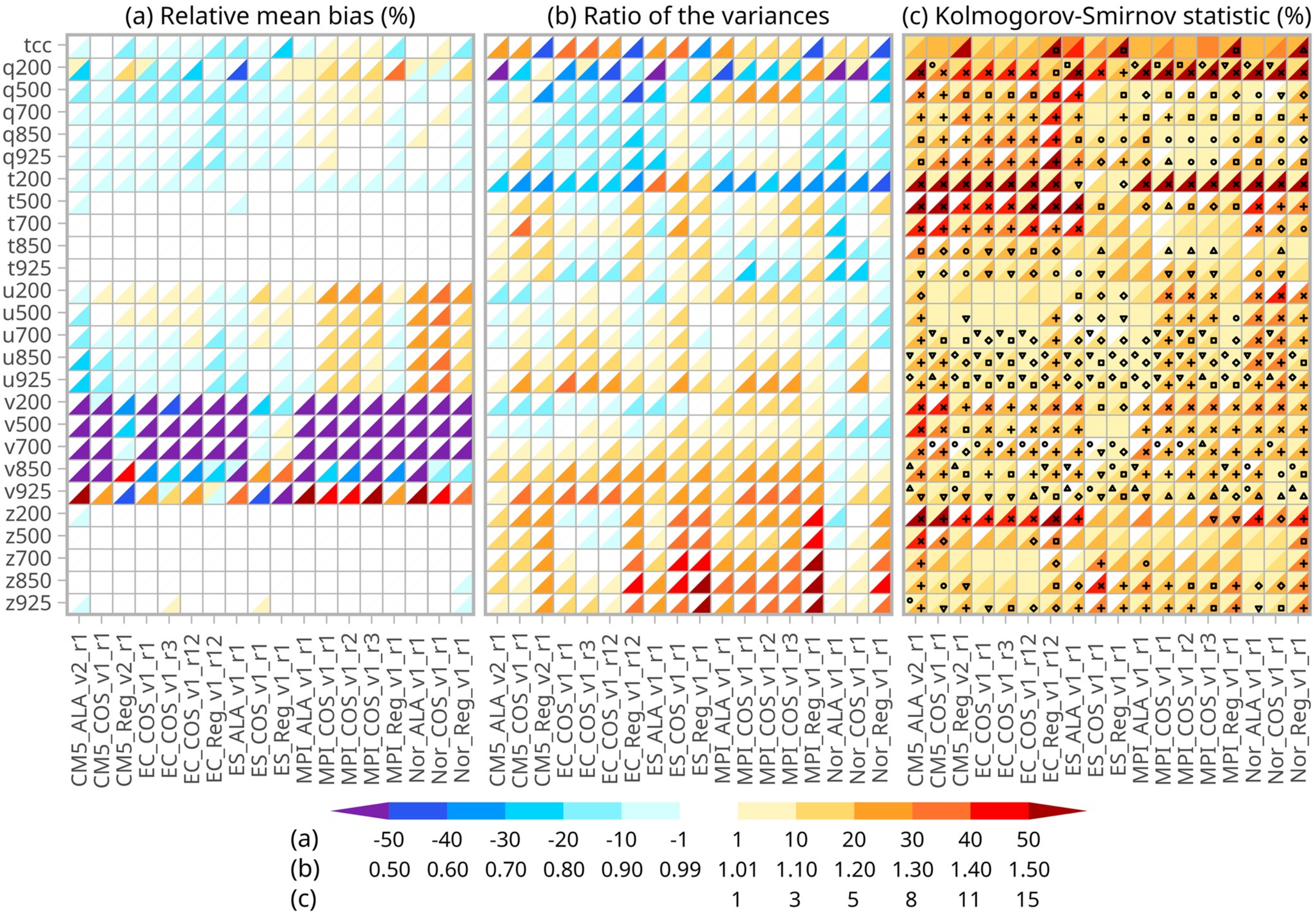

Figure 4. Portrait plots of both bias-corrected and raw EUR25 models against ERA5 arranged row-wise by predictor and column-wise by model. The metrics were calculated pixel-wise from daily values between 1979 and 2005 and the values shown correspond to the median of all the pixels. The predictors are named by variable and pressure level in hPa. The upper left (lower right) triangles are related to the bias-corrected (raw) predictors. Note that the color scale is shared, yet with different numerical breaks. To highlight variables whose performance is remarkably close to the perfect value (zero), a modification to the original portrait plots is introduced, that is, a class in white with values between −1% and 1% in (a), 0.99 and 1.01 in (b) and, 0% and 1% in (c). Additionally, the symbols in (c) are related to the number of pixels (from 1,024) in which the null hypothesis of the Kolmogorov–Smirnov test (both sets are drawn from the same continuous distribution), can be rejected (p-values < 0.05), that is, no symbol → 0, 0 < ∘ < 10, 10 ≤ △ < 20, 20 ≤ ▽ < 50, 50 ≤ ⋄ < 200, 200 ≤ □ < 500, 500 ≤ +< 1,024, and × → 1,024 (pixels).

of the bias-correction method for the predictor sets, which is out of the scope of the present project. Still, these techniques could introduce unexpected artifacts for the projected climate conditions, for example, under unseen climate (extrapolating conditions), and should be carefully evaluated with a similar approach to the one presented.

3.3. Historical Period Evaluation

Another crucial aspect for trustworthy PP statistically downscaled projected climate relies on the credibility of the downscaled predictands during the historical period (Maraun & Widmann, 2018). Since GCMs are usually "free-running" models, they are not synchronized with for example, the ERA5 reanalysis, thus, no strict temporal comparison should be done. Thus, the tests are carried out on a climatological basis, for which we selected the training period 1979–2005. The subfigures in Figure 5 were developed to elucidate relevant statistical aspects of the observed and downscaled predictands.

Figure 5 shows the daily PDFs of the observed values, EUR10 and the downscaled daily values for a selection of GRCMCs (for all the combinations of downscaled values see Figure A2 for Dresden). Since the deterministic approach uses the *expected value* of the correspondent fitted PDFs, which would generate more values closer to the mean and less toward the tails, the shown values correspond to the stochastical projections. The analogous EUR10 variables show notable differences against ReKIS (on dark gray) in their distribution for almost all vari-

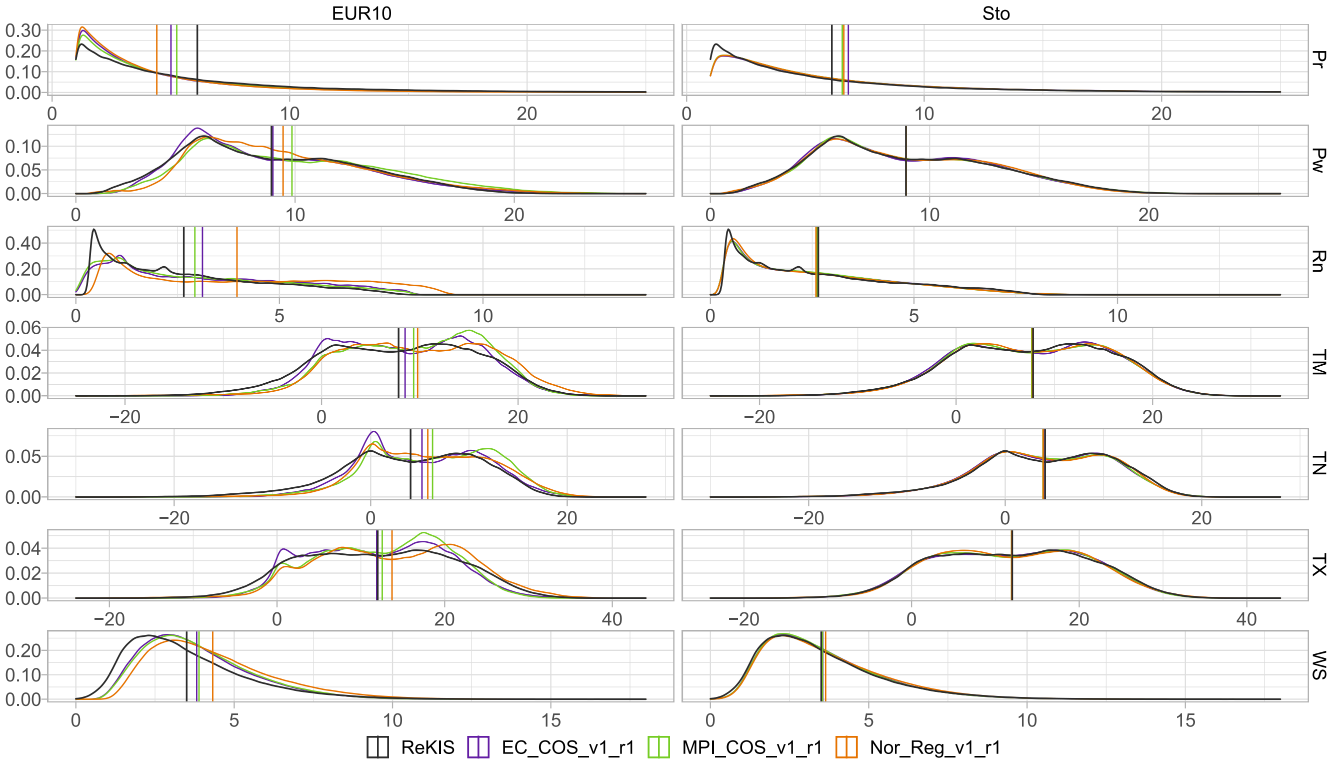

Figure 5. PDFs of daily values for selected GRCMCs for EUR10, the downscaled predictands (stochastic runs only), and the observed ones, between 1979 and 2005 (training period—historical runs for the GRCMCs), divided column-wise by data set and row-wise by predictand. Pr shows values ≥1 mm · day^{-1} only. The vertical lines show the mean value per data set.

ables, but especially for Rn, which agrees with the findings of Katragkou et al. (2015). On the other hand, it can be appreciated how most of the downscaled datasets behave remarkably similarly to the observed ones. The last two mentioned features of both datasets elucidate the added value of our approach by obtaining a more accurate distribution of daily values in comparison with RCM output. Naturally, there are some features of the observed predictands which are not entirely captured by the downscaling method, for example, (a) the values close to the wet-day threshold in Pr, and (b) the peak before 2.5 kW · h · m^{-2} in Rn, among others. In particular, (a) can be partly explained by the large range and weight on the tail for Pr, that is, between 0 and ~320 mm · day^{-1} (corresponding to the central European floods of August 2002), which forces a shift in the BG PDF and its mean values (vertical lines). The mean values of Pr show a systematic bias of approximately 1 mm · day^{-1} (absent for the validation data set), which needs to be taken into account for further uses of the data set.

Figure A1 (complement of Figure 5) displays the daily averages for the mean, and the 5th and 95th percentiles, to show the behavior of the extreme values. It can be observed again that Pr is the hardest predictand to properly model, with some downscaled models showing higher p95 precipitation values for boreal autumn, instead of summer (expected), for example, NorESM1-M_RegCM4-6_v2_r1 and NorESM1-M_ALADIN63_v1_r1 (in Figure A2) and Figure A3. Most of the presented stochastic results seem to have quite a good fit for the three metrics shown. Interesting biases, which are minor but consistent, can be seen for the stochastic values of Rn, for example, (a) the underestimation of the mean (during summer) and the fifth percentile (for most of the year) values, yet considerably less than EUR10, (b) the overestimation of the 95th percentile from July until the end of the year. These findings could be explained by the fact that tcc is the variable which obtained higher KSSs (see Figure 4c) conjugated with the previously mentioned sensitivity of Rn to it. The latter, besides lower radiation being the evident physical outcome of higher cloud coverage, was proven by mistake during the first iteration toward these results. The range of tcc in ERA5 is [0,1] and [0,100] (%) in EUR25, which was initially neglected and caused to have extremely noisy downscaled values for Rn (not shown) but also, though less severely, for Pr and TX. This anecdote converges with the second assumption ("informativeness") of the PP methodology, which states that the selected predictors should explain a large fraction of the variability. Additionally, the coupling of

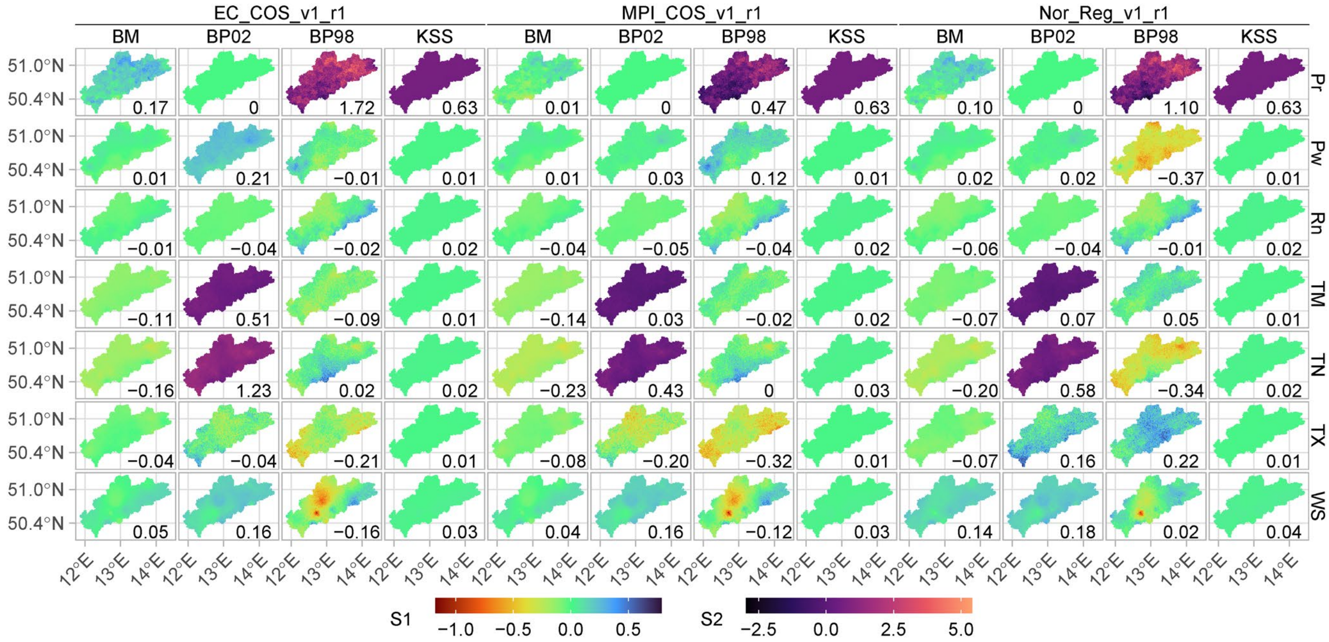

Figure 6. Maps of selected metrics and GRCMCs between the observed values and the downscaled ones for the training period (stochastic runs), arranged column-wise per metric, clustered by GRCMC, and row-wise per predictand. Pr shows values ≥1 mm · day^{-1} only. Note that, for enhanced visual clarity, scale 1 (S1) represents all the maps except for BP98 and KS for Pr, and BP02 for TM and TN, for which scale 2 (S2) applies. The value at the bottom right corner denotes the average value per GRCMC-predictand combination.

the dependency of Pr on tcc and the described differences among the PDFs for tcc in ERA5 and bias-corrected EUR25 (Figure 4c) could partly explain the systematic biases seen in Pr, particularly about the wet-day threshold.

Since Figures 5 and A1 show the aggregated daily data for all the study region, smaller scale particularities might be easily overlooked. Therefore, Figures A3, A4, and A5 were added to display the performance of the three representative subregions. Note that generally the downscaled data resembles the observed one better than the EUR10 output but it is even clearer under topographically complex terrain (Fichtelberg), as seen particularly for WS in Figure A4, which is severely underestimated in EUR10.

Figure 6 displays the maps of a selection of metrics and GRCMCs, aggregated for the training period for all the study region. It is noteworthy that most of the combinations which are represented by S1 indicate a robust performance, in this case, also on the spatial distribution aspect. The latter can be inferred by the distributed values close to zero, as well as their absolute means. Once more, it is shown that Pr remains the most challenging variable to accurately model, as seen in BP98 and KSS (represented by S2). The differences in BP98 are rather small for extreme precipitation events (−2.5 to 5 mm · day^{-1}). The shortcomings in KSS are important and could be explained by the aforementioned discrepancies close to the wet-day threshold. Still, after this threshold, the shape of the related downscaled PDFs resemble the observed ones satisfactorily. The differences observed in BP02 for TM and TN are relatively minor, except for EC_COS_v1_r1. Still, the biases for the aforementioned GRCMC (0.51 and 1.23°C, respectively) are also rather small for extreme events.

Moreover, it should be clarified that no hard limits or boundaries were set for the downscaled values in contrast to for example, Lange (2019), with which the stochastic extrapolated values could be substantially higher (lower) than the observed ones. Also, a side effect of the stochastic approach is the detriment of the spatial correlation (Quesada-Chacón et al., 2022), which should be addressed in future research. Despite the hereby argued shortcomings among the statistical properties of the pseudo-observations, the results are highly satisfactory. Thus, we consider the downscaled projections to be of considerable value for impact modelers.

3.4. Downscaled Projected Climate

Due to its higher amount of members from the EUR25 subset and current anthropogenic CO_2 usage, RCP85 will be employed hereafter for analysis. To provide impact modelers with a curated selection of best-performing downscaled models, a subset of VALUE metrics (KSS, BM and Bp98) were calculated for the GRCMCs to rank them according to the performance in the training period. The top nine models (out of 18) according to their performance between 1979 and 2005, mentioned by increasing rank with the abbreviated GRCMC nomenclature, are: EC_COS_v1_r1, MPI_COS_v1_r1, EC_COS_v1_r12, EC_COS_v1_r3, Nor_Reg_v1_r1, MPI_COS_v1_r3, Nor_COS_v1_r1, ES_COS_v1_r1, MPI_COS_v1_r2. Note that three runs each of both ICHEC-EC-EARTH and MPI-M-MPI-ESM-LR, all with CLMcom-ETH-COSMO-crCLIM-v1-1 (ranked #1, #3 and #4 and #2, #6 and #9, respectively), are between these top nine and possibly have a high degree of similarity and dependency.

As discussed, the biases during the historical period are rather close to zero, except for Pr. Due to these minor but present discrepancies during the historical period, it was decided to calculate the differences for each GRCMC per future subperiod individually against its historical period (delta change approach), instead of from ReKIS. Consequently, Figure 7 shows the spatial distribution of the aforementioned differences among predictands, GRCMCs (ranked #1, #2, and #5, respectively, to show more interdependent GRCMCs) and future periods.

The results in Figure 7 are generally in agreement for FF with studies on coarser resolution, for example, average temperature increases ∼3–4°C and increments of less than 1 mm $\cdot$ day^{-1} for SDII, for example, Baño Medina et al. (2022). As observed more clearly in variables like Pw, TM, TN, and TX, the downscaled values toward the end of the century are considerably higher than the ones during the observed period, as expected for the worst-case scenario. This fact relates to the last PP assumption: suitability of the transfer function structure, that is, "… the influence of the predictors on the predictand (possibly including interactions between the predictors) needs to be reasonably well incorporated. In a different climate, predictors and predictand will likely enter values that are outside the observed range. Therefore, for the downscaling model to be sensible, its structure needs to sensibly allow for at least moderate extrapolations…" (Maraun & Widmann, 2018). Thus, due to the shown results, we corroborate the suitability of the DL models employed. The conjunction of this third assumption with the second one (*informativeness*) is known as the "time-invariance or stationarity assumption: if all predictors relevant for climate change are included, and their influence on the predictand is sensibly modeled, also beyond observed states, the model is valid in a future climate" (Maraun & Widmann, 2018). Therefore, we consider to have complied, to a high degree, with all three major assumptions of the PP methodology, and thus, the models are valid under future conditions. Consequently, the related projections should be valid as well.

Moreover, note that the differences between EUR10 and the downscaled values are largely similar for all the predictands, both spatially distributed and averaged. The latter can be interpreted as properly conserving the climate change signal within the GRCMs. The largest discrepancies among the differences are seen for Rn, which are consistently lower than the downscaled values. This observation agrees with previously detected shortcomings of Rn in RCMs (Katragkou et al., 2015). As illustrated in for example, Figure 5, the downscaling methodology corrects the PDFs for historical conditions, while EUR10 exhibits significant deficiencies for that period, which are likely to persist in the future. Also, note how the downscaled values seem to improve substantially the spatial distribution of highly consistent differences, for example, (a) Pw in Figure 7a, (b) WS in Figure 7a, and, (c) Pr_SDII in Figure 7b, more clearly observed with MPI_COS_v1_r1. The achieved spatial refinement is what we aspired to and is a highly satisfying outcome which we believe further validates the aggregated value of the downscaling approach.

The projected possible unobserved future climatic conditions increase the possibilities of univariate extreme events, as well as compound extreme events, which need to be analyzed with daily data, as here presented. Yet, the analysis of compound extreme events is out of the scope of the present paper. Furthermore, the presented data set can be further analyzed to for example, calculate extreme climate indices (Zhang et al., 2011) to provide summarized numerical estimates of future changes. Still, the aforementioned indices were designed for precipitation and temperature only, calculated univariately. Due to the growing need for multivariate ensembles of projections, analogous multivariate indices are sorely needed.

Additionally, the trend of the climate change signals are shown in Figure A6, which displays a matrix of time series plots for the yearly averages of the entire downscaling region for all the combinations of predictands and GRCMCs. As expected from free-running GRCMCs, the observed and downscaled values for the same year can have appreciable differences, nonetheless, the yearly variability of the downscaled values resembles the one of ReKIS for the past and agrees with the signals observed in studies with a coarser resolution (e.g., Baño-Medina et al., 2022; Lange, 2019).

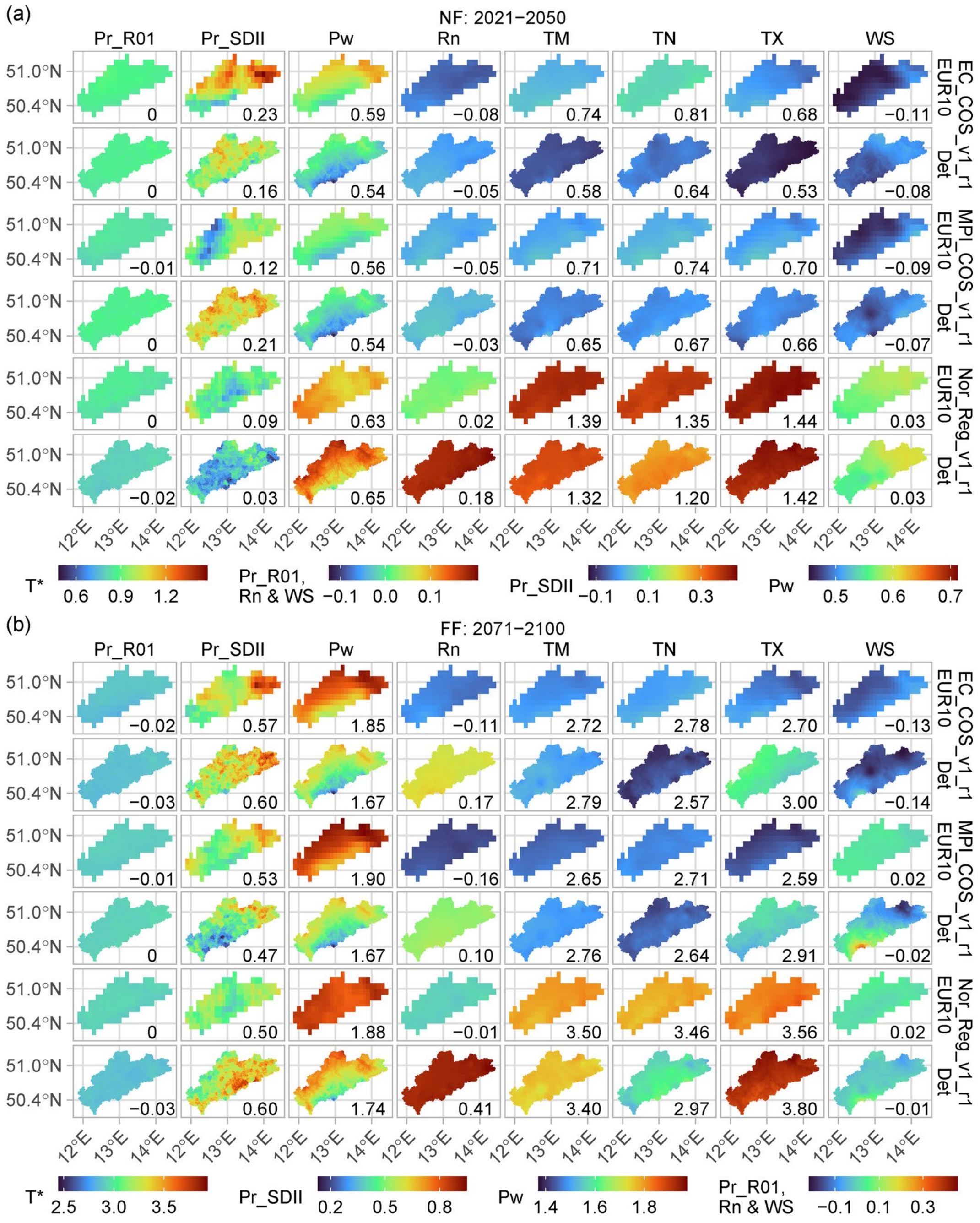

Figure 7. Maps of the differences between averaged climatologies of the future and the historical runs (delta change approach) for all predictands (column-wise) of a selection of GRCMCs. The different GRCMCs are ordered row-wise in ascending ranking order, note that both EUR10 and the deterministically downscaled values (Det) are shown for each GRCMC. Precipitation was separated into the fraction of wet days (R01) and SDII. The value at the bottom right corner denotes the absolute average per combination. For better appreciation of the spatial variability, the predictands were grouped by similar range in their differences for future scenarios. Note the shared numerical breaks in the color scales.

3.5. Caveats

Bearing transparency in mind and with the hope that future users will be able to make the most out of the hereby presented data set, we list the following caveats:

- Both stochastic and deterministic runs are provided for all the predictands. Its selection will depend on the application, for example, stochastic runs should be preferred if extreme values are relevant.

- Some of the yearly averaged values for CM5_COS_v1_r1 in 2065 showed relatively high values in 2065 (Pr_R01, Pw and TN), which might be explained by rather extraordinary atmospheric conditions for this year, causing systematic extrapolation difficulties.
- The consistent biases in Pr should be carefully taken into account in impact models, which we believe is highly influenced by the persistent biases within the corrected EUR25 predictors, particularly tcc. It might not be an issue for higher values (>10 mm $\cdot$ day^{-1}) but would for values close to the wet-day threshold. Additional bias-correction methods, aware of the extrapolation needs of the approach, could be employed to solve this issue.
- EC_COS_v1_r1 was collectively the highest ranked GRCMC. Nevertheless, it was the last ranked model (#18) for Pr, that is, most of the other predictands had outstanding performances with this GRCMC.
- Some particularities may arise in the case of stochastic runs:
  - Loss of the spatial correlation within each predictand.
  - The relationship among closely related characteristics of different predictands may be lost. This could also be the case for deterministic runs. This issue should be addressed with for example, the aforementioned approach of training the transfer functions with a multivariate copula loss function and/or introducing physical constraints within the DL models (Hess et al., 2022).
  - Unexpected artifacts, such as TM being lower than TN. This could be addressed with an approach similar to Lange (2019), where instead of modeling TN, and TX individually, the daily range and the skewness of the daily temperature cycle were modeled.

4. Summary and Outlook

In the present paper, we extended the pre-established methodology (Baño-Medina et al., 2022; Quesada-Chacón et al., 2022) to provide the highest statistically downscaled spatio-temporal resolution multivariate projection ensemble currently available. It is the first daily 1 km ensemble data set until the end of the century. For this, we employed deep learning models to develop transfer functions under the perfect prognosis statistical downscaling approach for: precipitation, water vapor pressure, radiation, wind speed, and, maximum, mean and minimum temperature. Such high spatio-temporal resolution multivariate ensembles are urgently needed for impact studies (Lange, 2019). The transfer functions are based on the *U-Net* and the *U-Net++* architectures, which have proven to substantially improve their performance. A subset of the ReKIS data set for Saxony was selected as predictand. The predictor set under perfect conditions is ERA5 with a total of 26 atmospheric variables. The projected ensemble is based on CORDEX EUR11, which includes 18 runs with RCP85, 4 with RCP26, and 1 with RCP45.

Several metrics from the VALUE framework were employed to quantify the performance of the transfer functions, which is, in general, highly satisfactory, also for the newly introduced predictands, that is, water vapor pressure, radiation, and wind speed, all for which the Gamma loss function yielded the best fit. The models for the three daily temperatures showed adequate results with the Gaussian loss function, and, as expected, precipitation remains the most challenging variable to accurately model, even with the Bernoulli-Gamma loss function.

Special consideration was given to the assumptions that support the perfect prognosis approach (Maraun & Widmann, 2018). Additional experiments and tests were executed to ensure, to a high degree, compliance with those major three assumptions, which were mostly fulfilled. The first measure consisted in processing the EUR11 ensemble to ensure its "realistic" and "bias-free" features, for which the Scaling Delta bias-correction method from Baño Medina et al. (2022) was employed. The EUR11 data set was upscaled (EUR25) to match the grid of ERA5, 0.25°, and then bias-corrected. Despite the considerable efforts, minor discrepancies between the statistical characteristics of some predictors were observed, for example, PDF differences between ERA5 and EUR25 for total cloud fraction, which could be responsible for the slight performance drops and systematic biases observed in the projected ensemble, particularly for precipitation and radiation. Nevertheless, thorough conformity with the other two assumptions (*time-invariance*) appears to be less debatable, thus, we consider the transfer functions to be highly suitable for downscaling the ensemble. We suggest analyzing further bias-correction methods for the predictors to address the aforementioned issues, for example, a quantile mapping approach aware of the expected unobserved atmospheric states.

In general, the averaged downscaled projected climate agrees with coarser estimates in predictands like precipitation and temperature, that is, increases of ~3–4°C in temperature and less than 1 mm $\cdot$ day^{-1} in wet-days toward the end of the century. Additionally, the downscaled climate change signal generally agrees with the corresponding CORDEX one, except for Rn, which shows biases confirmed in the literature (Katragkou et al., 2015). The spatial features from CORDEX were satisfactorily refined with the downscaling approach. The provided daily data can provide better insights into extreme events than other temporarily aggregated data available. We hope that this data set proves useful to impact modelers interested in the region since it could drive more precise and diverse models in fields like agriculture, ecology, and flood risk. We also provide a list of caveats for the potential users and enumerate the best-performing combinations of GCM-RCMs.

Although the study area is very localized, the shown methodology is scalable to other regions, datasets, and spatio-temporal resolutions. Generally, the methodology would profit from higher spatio-temporal resolution predictors and predictands. One straightforward possible extension of the shown workflow would be to use recent daily 1 km datasets for the past (Karger et al., 2021, 2022) as predictands (with ERA5 and CORDEX as predictors) in similar applications for anywhere in the world. Additionally, the latest generation available of GCMs, CMIP6 (Eyring et al., 2016), could be also downscaled with similar approaches.

Major improvements in the present methodology could be achieved by employing multivariate copulas (e.g., François et al., 2020) as the loss function of the deep learning models. This would ideally allow better preservation of the spatio-temporal features, and the interdependencies (coherence) between near-surface variables under both stochastic and deterministic conditions. Such a data set would allow better analysis and prognosis of future compound events (Zscheischler et al., 2018), which is currently limited due to the univariate nature of the employed methodology. To properly analyze such features, comprehensive multivariate metrics frameworks analogous to VALUE (Gutiérrez et al., 2019; Maraun et al., 2014) are needed. Finally, the integration of: (a) enhanced predictor bias-correction methods, (b) physical constraints within the deep learning models (e.g., Hess et al., 2022), (c) mechanistic statistical downscaling features (e.g., CHELSA, Karger et al., 2017), as well as (d) the aforementioned multivariate coherence, could prove to be the optimal next iteration of the perfect prognosis statistical downscaling methodology.

Appendix A: Supplementary Material

This section comprises a collection of high-resolution figures that serve as valuable complements to the visual representations included in the main body of the paper. By offering additional views and detailed illustrations,

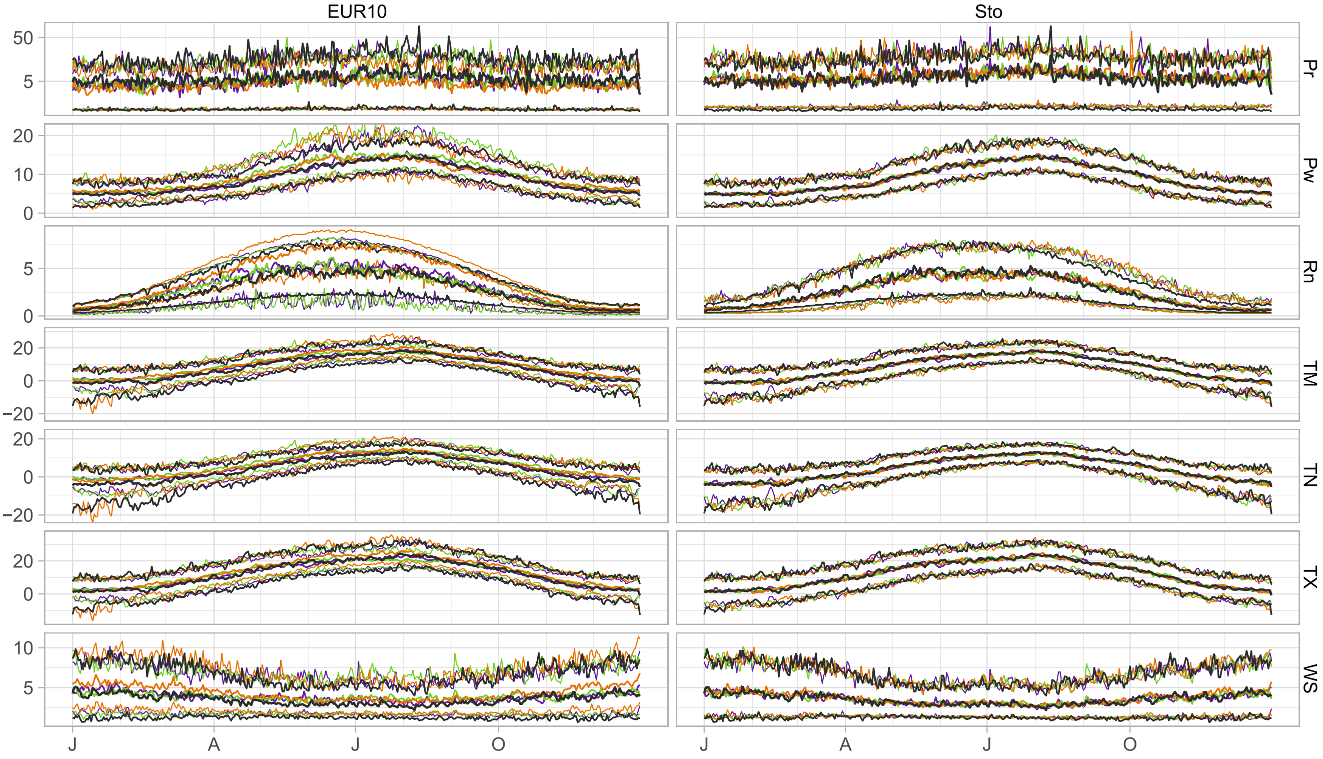

Figure A1. Day of the year plots of a selection of GRCMCs for EUR10, the downscaled predictands (stochastic runs only), and the observed ones, between 1979 and 2005, divided column-wise by data set and row-wise by predictand. Pr shows values ≥1 mm · day^{-1} only. Each panels shows the 5th and 95th percentiles, and the mean. Each vertical grid line corresponds to the month's first day. To improve the visualization of Pr, the *y-axis* has a square-root transformation. Complement of Figure 5.

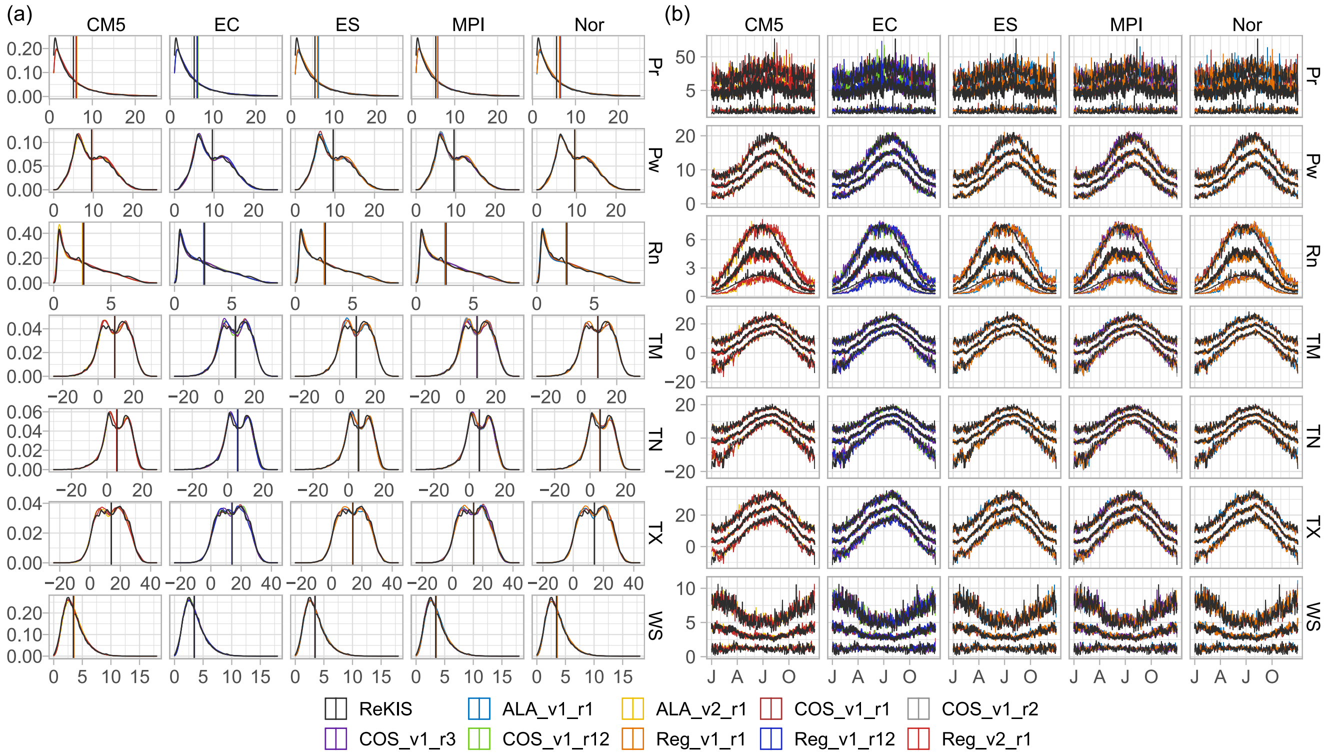

Figure A2. Statistical summaries for the subregion of Dresden (see Figure 1a) of all the downscaled predictands, and the observed, between 1979 and 2005 (training period—historical runs for the GRCMCs), divided column-wise by GCM and row-wise by predictand, the colors indicate either the observed data or the RCMs; stochastic runs only. (a) PDFs of daily values, the vertical lines show the mean value per data set. (b) Day of the year plots for the 5th and 95th percentiles, and the mean. Each vertical grid line in (b) corresponds to the month's first day. To improve the visualization of Pr, the *y-axis* in (b) has a square-root transformation.

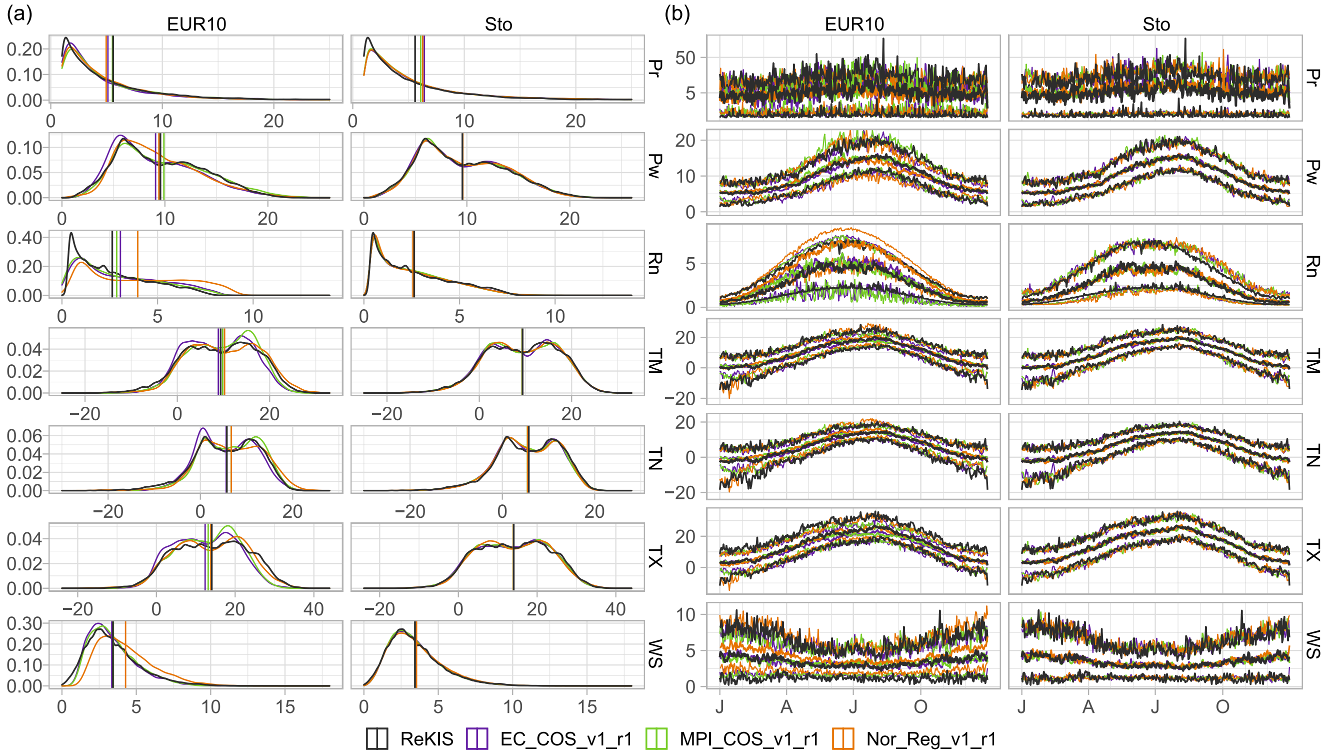

Figure A3. Statistical summaries for the subregion of Dresden. (a) Is analog to Figure 5 and (b) to Figure A1.

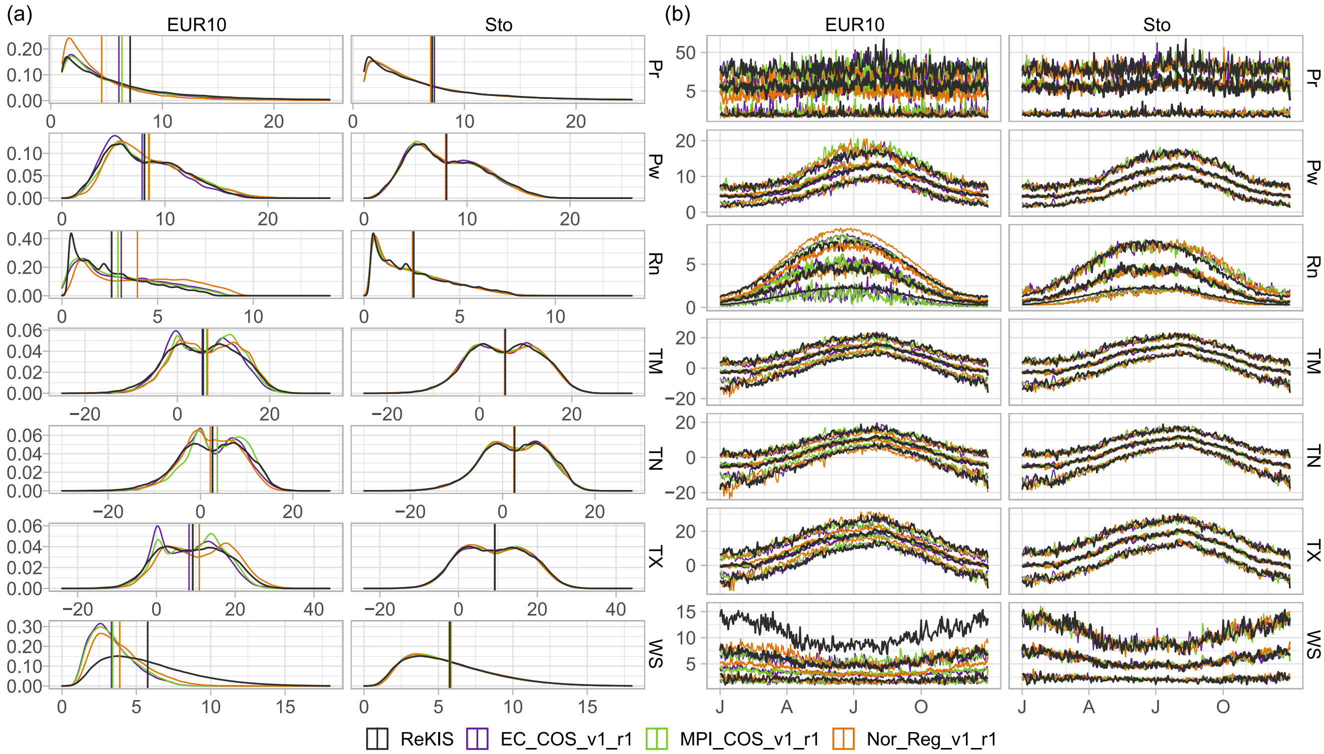

Figure A4. Same as Figure A3 but for Fichtelberg.

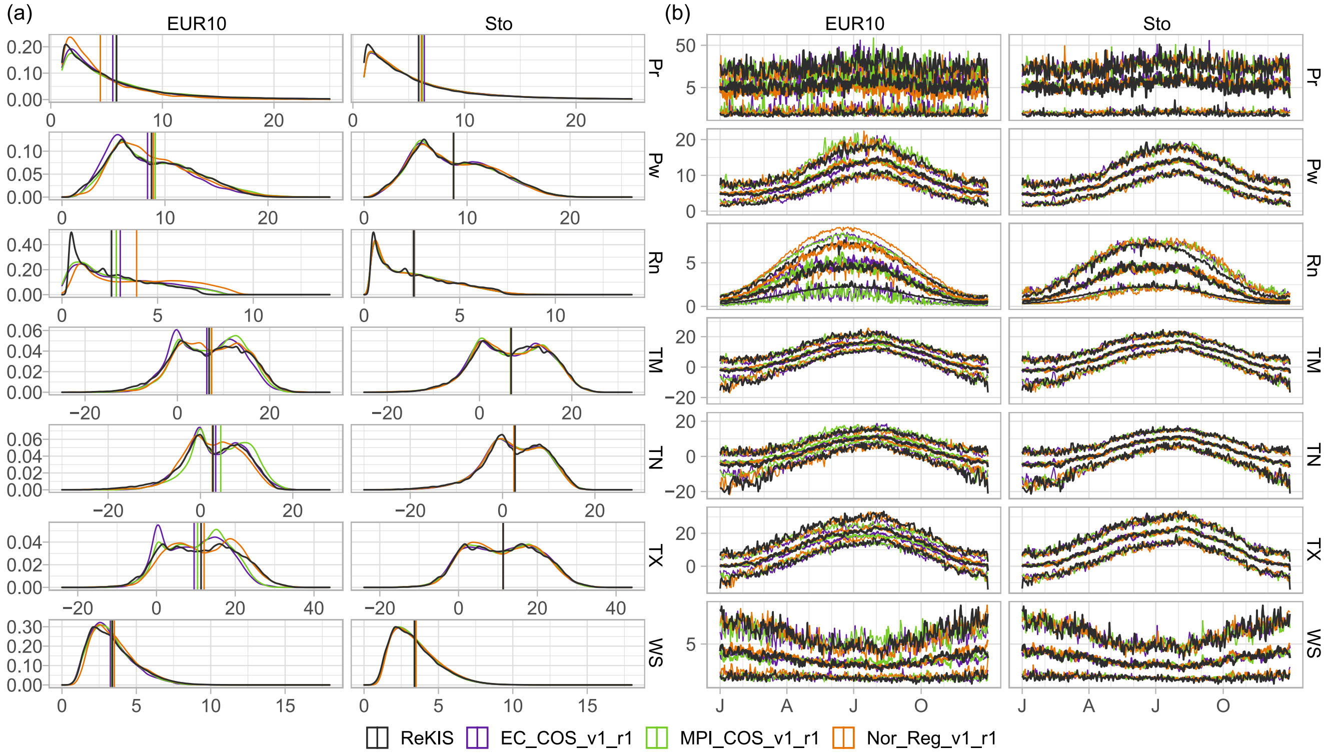

Figure A5. Same as Figure A3 but for Vogtland.

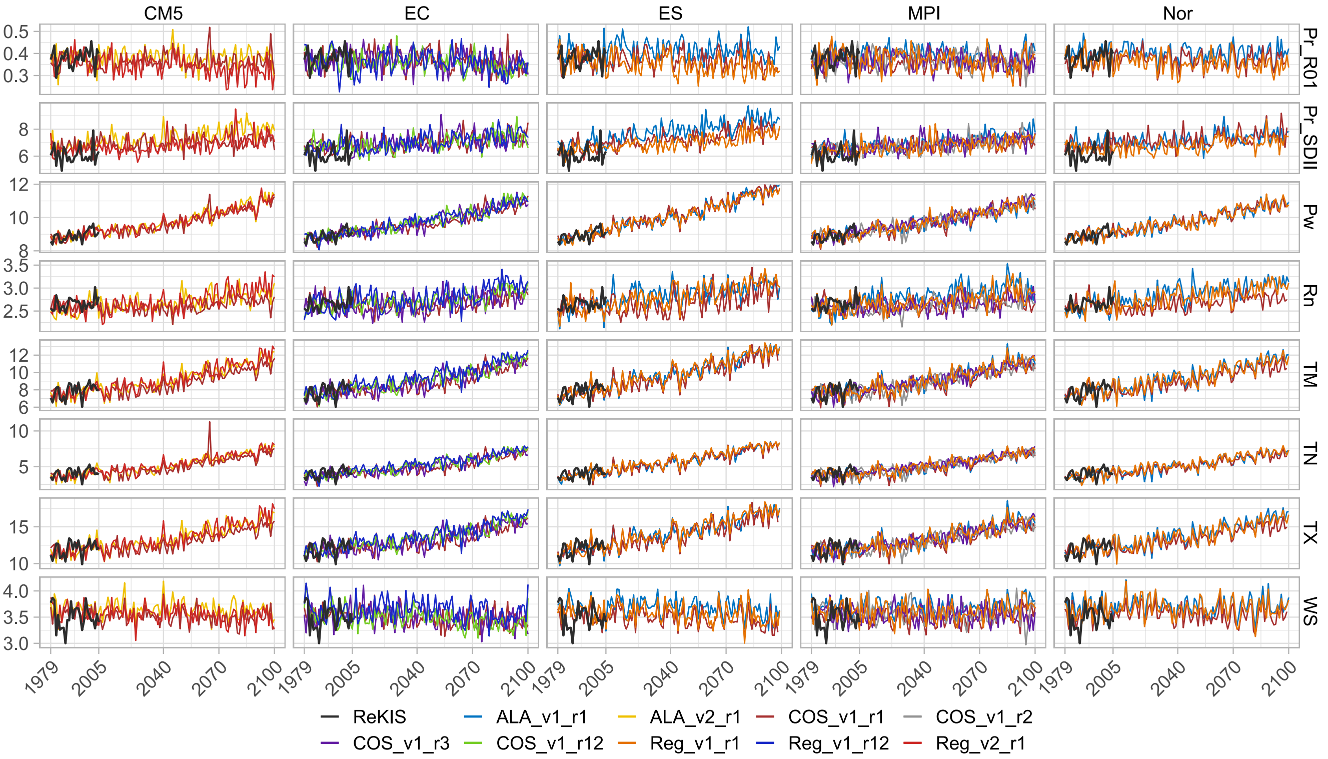

Figure A6. Yearly averaged time series for the deterministic projections of all the predictands and ensemble members from 1979 until the end of the century (RCP85). The whole study area was used for the calculations. Note that precipitation was separated into the fraction of wet days (R01) and SDII.

these supplementary figures enhance the overall understanding of the research, enriching the methodology and supporting the results presented in the study. Each figure is sequentially numbered to correspond with its reference in the main text, facilitating seamless cross-referencing for readers.

Acknowledgments
We want to thank: the European Centre for Medium-Range Weather Forecasts (ECMWF) for providing the ERA5 reanalysis data, the Regionales Klimainformationssystem (ReKIS) for the gridded observed data, and the Coordinated Regional Climate Downscaling Experiment (CORDEX) for the dynamically downscaled projections. We would also like to thank Zenodo for providing repositories with larger data quotas to deposit our datasets. We appreciate the generous allocations of computational resources by the Center for Information Services and High Performance Computing (ZIH) of the Technische Universität Dresden and the support of the Competence Center for Scalable Data Services and Solutions Dresden/Leipzig (ScaDS). This research is funded by the European Social Fund and co-financed by tax funds based on the budget approved by the members of the Saxon State Parliament (100380876). We express our gratitude to the anonymous referees for their valuable feedback, which enhanced the quality of this paper. Open Access funding enabled and organized by Projekt DEAL.

Data Availability Statement

The raw datasets and their respective sources are: ERA5 (Hersbach et al., 2020), CORDEX-EUR11 (Jacob et al., 2014), and ReKIS (2021). The pre-processed predictors and predictands are available at Quesada-Chacón (2023c). Due to its size, only a subset of the complete ensemble could be hosted at Quesada-Chacón (2023b). The subset contains the downscaled values of all predictands for different high-performing combinations of GCM-RCMs of both stochastic and deterministic runs for eight historical runs, eight RCP85 runs and one RCP26 run. The other ensemble members can be obtained through the corresponding author. The code used to train the DL models, bias-correct the predictors and to generate the projections can be found at Quesada-Chacón (2023a). The employed updated version (2.0.0) of the containerized software environment is hosted at Quesada-Chacón (2023d). The figures presented in this paper can be found in their original resolution at Quesada-Chacón (2023e), ensuring clear and high-quality visual representations for readers.

References

Baño-Medina, J., Manzanas, R., Cimadevilla, E., Fernández, J., González-Abad, J., Cofiño, A. S., & Gutiérrez, J. M. (2022). Downscaling multi-model climate projection ensembles with deep learning (DeepESD): Contribution to CORDEX EUR-44. *Geoscientific Model Development*, *15*(17), 6747–6758. https://doi.org/10.5194/gmd-15-6747-2022

Baño-Medina, J., Manzanas, R., & Gutierrez, J. M. (2020). Configuration and intercomparison of deep learning neural models for statistical downscaling. *Geoscientific Model Development*, *13*(4), 2109–2124. https://doi.org/10.5194/gmd-13-2109-2020

Bevacqua, E., De Michele, C., Manning, C., Couasnon, A., Ribeiro, A. F. S., Ramos, A. M., et al. (2021). Guidelines for studying diverse types of compound weather and climate events. *Earth's Future*, *9*(11), e2021EF002340. https://doi.org/10.1029/2021EF002340

Brun, P., Zimmermann, N. E., Hari, C., Pellissier, L., & Karger, D. N. (2022). Global climate-related predictors at kilometre resolution for the past and future. *Earth System Science Data Discussions*, *2022*, 1–44. https://doi.org/10.5194/essd-2022-212

Cannon, A. J. (2018). Multivariate quantile mapping bias correction: An N-dimensional probability density function transform for climate model simulations of multiple variables. *Climate Dynamics*, *50*(1), 31–49. https://doi.org/10.1007/s00382-017-3580-6

CORDEX. (2021). CORDEX - ESGF data availability overview. Retrieved from http://is-enes-data.github.io/CORDEX_status.html

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the coupled model intercomparison project phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

François, B., Thao, S., & Vrac, M. (2021). Adjusting spatial dependence of climate model outputs with cycle-consistent adversarial networks. *Climate Dynamics*, *57*(11), 3323–3353. https://doi.org/10.1007/s00382-021-05869-8

François, B., Vrac, M., Cannon, A. J., Robin, Y., & Allard, D. (2020). Multivariate bias corrections of climate simulations: Which benefits for which losses? *Earth System Dynamics*, *11*(2), 537–562. https://doi.org/10.5194/esd-11-537-2020

Gleckler, P. J., Taylor, K. E., & Doutriaux, C. (2008). Performance metrics for climate models. *Journal of Geophysical Research*, *113*(D6), D06104. https://doi.org/10.1029/2007JD008972

Gutiérrez, J. M., Maraun, D., Widmann, M., Huth, R., Hertig, E., Benestad, R., et al. (2019). An intercomparison of a large ensemble of statistical downscaling methods over Europe: Results from the VALUE perfect predictor cross-validation experiment. *International Journal of Climatology*, *39*(9), 3750–3785. https://doi.org/10.1002/joc.5462

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Hertig, E., Jahn, S., & Kaspar-Ott, I. (2023). Future local ground-level ozone in the European area from statistical downscaling projections considering climate and emission changes. *Earth's Future*, *11*(2), e2022EF003317. https://doi.org/10.1029/2022EF003317

Hess, P., Drüke, M., Petri, S., Strnad, F. M., & Boers, N. (2022). Physically constrained generative adversarial networks for improving precipitation fields from Earth system models. *Nature Machine Intelligence*, *4*(10), 828–839. https://doi.org/10.1038/s42256-022-00540-1

Höhlein, K., Kern, M., Hewson, T., & Westermann, R. (2020). A comparative study of convolutional neural network models for wind field downscaling. *Meteorological Applications*, *27*(6), 1–31. https://doi.org/10.1002/met.1961

Huang, J. (2018). A simple accurate formula for calculating saturation vapor pressure of water and ice. *Journal of Applied Meteorology and Climatology*, *57*(6), 1265–1272. https://doi.org/10.1175/JAMC-D-17-0334.1

Huth, R. (2005). Downscaling of humidity variables: A search for suitable predictors and predictands. *International Journal of Climatology*, *25*(2), 243–250. https://doi.org/10.1002/joc.1122

IPCC. (2021). In V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, et al. (Eds.), *Climate change 2021: The physical science basis. Contribution of Working Group I to the sixth assessment report of the Intergovernmental Panel on Climate Change*. Cambridge University Press. https://doi.org/10.1017/9781009157896

Jacob, D., Petersen, J., Eggert, B., Alias, A., Christensen, O. B., Bouwer, L. M., et al. (2014). EURO-CORDEX: New high-resolution climate change projections for European impact research. *Regional Environmental Change*, *14*(2), 563–578. https://doi.org/10.1007/S10113-013-0499-2/

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., et al. (2017). Climatologies at high resolution for the Earth's land surface areas. *Scientific Data*, *4*, 1–20. https://doi.org/10.1038/sdata.2017.122

Karger, D. N., Lange, S., Hari, C., Reyer, C. P. O., & Zimmermann, N. E. (2022). CHELSA-W5E5 v1.0: W5E5 v1.0 downscaled with CHELSA v2.0. https://doi.org/10.48364/ISIMIP.836809.3

Karger, D. N., Schmatz, D. R., Dettling, G., & Zimmermann, N. E. (2020). High-resolution monthly precipitation and temperature time series from 2006 to 2100. *Scientific Data*, *7*(1), 1–10. https://doi.org/10.1038/s41597-020-00587-y

Karger, D. N., Wilson, A. M., Mahony, C., Zimmermann, N. E., & Jetz, W. (2021). Global daily 1 km land surface precipitation based on cloud cover-informed downscaling. *Scientific Data*, *8*(1), 1–18. https://doi.org/10.1038/s41597-021-01084-6

Katragkou, E., García-Díez, M., Vautard, R., Sobolowski, S., Zanis, P., Alexandri, G., et al. (2015). Regional climate hindcast simulations within EURO-CORDEX: Evaluation of a WRF multi-physics ensemble. *Geoscientific Model Development*, *8*(3), 603–618. https://doi.org/10.5194/gmd-8-603-2015

Körner, P., Vorobevskii, I., Kronenberg, R., & Homoudi, A. (2022). *Erzeugung eines lückenlosen, stationsbasierten und rasterbasierten Klima-Referenzdatensatzes für Sachsen für den Zeitraum 1961 bis 2020* (Tech. Rep.). Sächsisches Landesamt für Umwelt, Landwirtschaft und Geologie (LfULG). Retrieved from https://rekisviewer.hydro.tu-dresden.de/viewer/rekis_domain/KlimRefDS_Dokumentation/Klimatologische_Datengrundlagen_Modul_I__LfLULG_Schriftenreihe_18_2022__07_2022.pdf

Lange, S. (2019). Trend-preserving bias adjustment and statistical downscaling with ISIMIP3BASD (v1.0). *Geoscientific Model Development*, *12*(7), 3055–3070. https://doi.org/10.5194/gmd-12-3055-2019

Li, X., Li, Z., Huang, W., & Zhou, P. (2020). Performance of statistical and machine learning ensembles for daily temperature downscaling. *Theoretical and Applied Climatology*, *140*(1–2), 571–588. https://doi.org/10.1007/s00704-020-03098-3

Maraun, D., & Widmann, M. (2018). *Statistical downscaling and bias correction for climate research*. Cambridge University Press. https://doi.org/10.1017/9781107588783

Maraun, D., Widmann, M., Gutiérrez, J. M., Kotlarski, S., Chandler, R. E., Hertig, E., et al. (2014). VALUE: A framework to validate downscaling approaches for climate change studies. *Earth's Future*, *3*, 1–14. https://doi.org/10.1002/2014EF000259

Olmo, M. E., Balmaceda-Huarte, R., & Bettolli, M. L. (2022). Multi-model ensemble of statistically downscaled GCMs over southeastern South America: Historical evaluation and future projections of daily precipitation with focus on extremes. *Climate Dynamics*, *59*(9–10), 3051–3068. https://doi.org/10.1007/s00382-022-06236-x

Pierce, D. W., & Cayan, D. R. (2016). Downscaling humidity with Localized Constructed Analogs (LOCA) over the conterminous United States. *Climate Dynamics*, *47*(1–2), 411–431. https://doi.org/10.1007/s00382-015-2845-1

Quesada-Chacón, D. (2023a). dquesadacr/Downscaling_CORDEX: Submission to Earth's Future [Software]. Zenodo. https://doi.org/10.5281/zenodo.7570247

Quesada-Chacón, D. (2023b). Multivariate projected ensemble of "Downscaling CORDEX through deep learning to daily 1 km multivariate ensemble in complex terrain" [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.7559173

Quesada-Chacón, D. (2023c). Predictors and predictands for "Downscaling CORDEX through deep learning to daily 1 km multivariate ensemble in complex terrain" [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.7558945

Quesada-Chacón, D. (2023d). Singularity container for "Repeatable high-resolution statistical downscaling through deep learning" [Software]. Zenodo. https://doi.org/10.5281/zenodo.8059248

Quesada-Chacón, D. (2023e). High quality figures of "downscaling CORDEX through deep learning to daily 1 km multivariate ensemble in complex terrain". Zenodo. https://doi.org/10.5281/zenodo.8198925

Quesada-Chacón, D., Barfus, K., & Bernhofer, C. (2020). Climate change projections and extremes for Costa Rica using tailored predictors from CORDEX model output through statistical downscaling with artificial neural networks. *International Journal of Climatology*, *41*(1), 211–232. https://doi.org/10.1002/joc.6616

Quesada-Chacón, D., Barfus, K., & Bernhofer, C. (2022). Repeatable high-resolution statistical downscaling through deep learning. *Geoscientific Model Development*, *15*(19), 7353–7370. https://doi.org/10.5194/gmd-15-7353-2022

Quintero, F., Villarini, G., Prein, A. F., Krajewski, W. F., & Zhang, W. (2022). On the role of atmospheric simulations horizontal grid spacing for flood modeling. *Climate Dynamics*, *59*(11–12), 3167–3174. https://doi.org/10.1007/s00382-022-06233-0

Ramon, J., Lledó, L., Bretonnière, P.-A., Samsó, M., & Doblas-Reyes, F. J. (2021). A perfect prognosis downscaling methodology for seasonal prediction of local-scale wind speeds. *Environmental Research Letters*, *16*(5), 054010. https://doi.org/10.1088/1748-9326/abe491

ReKIS. (2021). Regionales klimainformationssystem sachsen, sachsen-anhalt, thüringen. Retrieved from https://rekis.hydro.tu-dresden.de

Rivington, M., Miller, D., Matthews, K. B., Russell, G., Bellocchi, G., & Buchan, K. (2008). Downscaling regional climate model estimates of daily precipitation, temperature and solar radiation data. *Climate Research*, *35*(3), 181–202. https://doi.org/10.3354/cr00705

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. CoRR, abs/1505.04597. Retrieved from http://arxiv.org/abs/1505.04597

San-Martín, D., Manzanas, R., Brands, S., Herrera, S., & Gutiérrez, J. M. (2017). Reassessing model uncertainty for regional projections of precipitation with an ensemble of statistical downscaling methods. *Journal of Climate*, *30*(1), 203–223. https://doi.org/10.1175/JCLI-D-16-0366.1

Schär, C., Fuhrer, O., Arteaga, A., Ban, N., Charpilloz, C., Girolamo, S. D., et al. (2020). Kilometer-scale climate models: Prospects and challenges. *Bulletin of the American Meteorological Society*, *101*(5), E567–E587. https://doi.org/10.1175/BAMS-D-18-0167.1

Schulzweida, U. (2021). CDO user guide. https://doi.org/10.5281/zenodo.5614769

Sørland, S. L., Schär, C., Lüthi, D., & Kjellström, E. (2018). Bias patterns and climate change signals in GCM-RCM model chains. *Environmental Research Letters*, *13*(7), 074017. https://doi.org/10.1088/1748-9326/aacc77

Taylor, K., Stouffer, R., & Meehl, G. (2012). An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, *93*(4), 485–498. https://doi.org/10.1175/BAMS-D-11-00094.1

Vandal, T., Kodra, E., Ganguly, S., Michaelis, A., Nemani, R., & Ganguly, A. R. (2018). Generating high resolution climate change projections through single image super-resolution: An abridged version. *Proceedings of the Twenty-Seventh Conference on Artificial Intelligence* (pp. 5389–5393). International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2018/759

Vrac, M. (2018). Multivariate bias adjustment of high-dimensional climate simulations: The rank resampling for distributions and dependences (R^2D^2) bias correction. *Hydrology and Earth System Sciences*, *22*(6), 3175–3196. https://doi.org/10.5194/hess-22-3175-2018

Wise, E. K. (2009). Climate-based sensitivity of air quality to climate change scenarios for the southwestern United States: Sensitivity of U.S. southwest air quality to climate change. *International Journal of Climatology*, *29*(1), 87–97. https://doi.org/10.1002/joc.1713

Zhang, X., Alexander, L., Hegerl, G., Jones, P., Tank, A., Peterson, T., et al. (2011). Indices for monitoring changes in extremes based on daily temperature and precipitation data. *Wiley Interdisciplinary Reviews: Climate Change*, *2*(6), 851–870. https://doi.org/10.1002/wcc.147

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. In *Lecture notes in computer science (including subseries Lecture notes in Artificial intelligence and lecture notes in bioinformatics), 11045 LNCS* (pp. 3–11). https://doi.org/10.1007/978-3-030-00889-5_1

Zscheischler, J., Fischer, E. M., & Lange, S. (2019). The effect of univariate bias adjustment on multivariate hazard estimates. *Earth System Dynamics*, *10*(1), 31–43. https://doi.org/10.5194/esd-10-31-2019

Zscheischler, J., Martius, O., Westra, S., Bevacqua, E., Raymond, C., Horton, R. M., et al. (2020). A typology of compound weather and climate events. *Nature Reviews Earth & Environment*, *1*(7), 333–347. https://doi.org/10.1038/s43017-020-0060-z

Zscheischler, J., Westra, S., Van Den Hurk, B. J., Seneviratne, S. I., Ward, P. J., Pitman, A., et al. (2018). Future climate risk from compound events. *Nature Climate Change*, *8*(6), 469–477. https://doi.org/10.1038/s41558-018-0156-3