



On the suitability of deep convolutional neural networks for continental-wide downscaling of climate change projections

Jorge Baño-Medina¹ · Rodrigo Manzanas² · José Manuel Gutiérrez¹

Received: 27 November 2020 / Accepted: 13 June 2021
© The Author(s) 2021

Abstract

In a recent paper, Baño-Medina et al. (Configuration and Intercomparison of deep learning neural models for statistical downscaling. preprint, 2019) assessed the suitability of deep convolutional neural networks (CNNs) for downscaling of temperature and precipitation over Europe using large-scale ‘perfect’ reanalysis predictors. They compared the results provided by CNNs with those obtained from a set of standard methods which have been traditionally used for downscaling purposes (linear and generalized linear models), concluding that CNNs are well suited for continental-wide applications. That analysis is extended here by assessing the suitability of CNNs for downscaling future climate change projections using Global Climate Model (GCM) outputs as predictors. This is particularly relevant for this type of “black-box” models, whose results cannot be easily explained based on physical reasons and could potentially lead to implausible downscaled projections due to uncontrolled extrapolation artifacts. Based on this premise, we analyze in this work the two key assumptions that are made in *perfect prognosis* downscaling: (1) the predictors chosen to build the statistical model should be well reproduced by GCMs and (2) the statistical model should be able to reliably extrapolate out of sample (climate change) conditions. As a first step to test the suitability of these models, the latter assumption is assessed here by analyzing how the CNNs affect the raw GCM climate change signal (defined as the difference, or delta, between future and historical climate). Our results show that, as compared to well-established generalized linear models (GLMs), CNNs yield smaller departures from the raw GCM outputs for the end of century, resulting in more plausible downscaling results for climate change applications. Moreover, as a consequence of the automatic treatment of spatial features, CNNs are also found to provide more spatially homogeneous downscaled patterns than GLMs.

Keywords Statistical downscaling · Regional climate change scenarios · Deep learning · Convolutional neural networks (CNNs) · Generalized linear models (GLMs)

1 Introduction

As a result of several decades of intense research, a variety of statistical downscaling approaches and techniques are nowadays available to fill the gap between the coarse resolution outputs provided by Global Climate Models (GCMs) and the local or regional information required for impact

studies—typically at individual locations or over regional interpolated grids covering the area of study—(Maraun and Widmann 2018). The local *predictand(s)* of interest (e.g. precipitation and temperature) are inferred from GCM outputs using statistical models which build on a set of large-scale atmospheric variables (*predictors* such as geopotential, temperature or humidity) explaining a large fraction of the local climate variability. Under the *perfect prognosis* (PP) approach, the different statistical downscaling models (hereafter SDMs) available (linear and generalized linear models, analogs, machine learning techniques) are trained based on historical observations for both predictands and predictors (the latter coming from a reanalysis). A number of intercomparison studies have reported the relative merits and limitation of state-of-the-art SDMs when ‘perfect’ reanalysis predictors are used, based on cross-validation. In particular,

✉ Jorge Baño-Medina
bmedina@ifca.unican.es

¹ Meteorology Group. Institute of Physics of Cantabria (IFCA), CSIC-University of Cantabria, Santander 39005, Spain

² Meteorology Group. Dpto. de Matemática Aplicada y Ciencias de la Computación, Universidad de Cantabria, Santander 39005, Spain

the VALUE initiative (pan-European network to validate downscaling methods, Maraun et al 2015) conducted the largest-to-date intercomparison with over 50 contributing SDMs (*VALUE Experiment 1*, Gutiérrez et al 2019).

In a recent paper, Baño-Medina et al (2019) assessed the suitability of deep convolutional neural networks (CNNs) for the downscaling of temperature and precipitation over Europe using large-scale ‘perfect’ reanalysis predictors. To do so, they used the experimental framework defined in VALUE (Experiment 1) to compare the results provided by CNNs with those obtained from a set of other more classical, standard techniques i.e., generalized linear models, concluding that CNNs are well suited for continental-wide applications. There have been similar studies over North America (Pan et al 2019) and China (Sun and Lan 2020), all showing that CNNs achieve similar or better performance than standard SDMs. Moreover, CNNs circumvent the problem of feature selection/extraction—which is highly case-dependent and becomes a very complex task to accomplish in classical downscaling methods—by performing an implicit manipulation of the input space in the internal structure of the network (Baño-Medina 2020).

However, these previous studies do not test the suitability and potential limitations of CNNs for climate change applications. This is particularly relevant due to the limited interpretability offered by these “black-box” models (Baño-Medina 2020) that may hinder extrapolation analysis. As a result, users are reluctant to use this kind of techniques and standard SDMs are still the preferred choice to downscale climate change scenarios (Gutiérrez et al 2019). Moreover, only a few studies have focused on the impact that the two key assumptions of *perfect prognosis* downscaling may have for climate change applications: (1) the predictors should be well reproduced by GCMs and (2) the statistical model should be able to generalize and extrapolate out-of-sample (e.g. climate change) conditions. This is crucial to assess the credibility of future climate information and avoid misadaptation (Pryor and Schoof 2020). Furthermore, these studies are mostly available at national or subnational levels (see, e.g., Gutiérrez et al 2013; San-Martín et al 2016; Manzananas et al 2015, 2020). At a continental level, there is an ongoing experiment aimed at analyzing these assumptions in the framework of the VALUE European initiative—now encompassed in the EURO-CORDEX collaboration (Jacob et al 2020),—but results are not available yet. Details of the experimental framework (*VALUE Experiment 2*, which we use in this study) are available at <http://www.value-cost.eu/validation>.

The present work is a first step to fill this knowledge gap by testing for the first time the suitability of CNNs for downscaling future climate change projections, extending the analysis done in Baño-Medina et al (2019) in the framework of VALUE (Experiment 2). We analyze the two

aforementioned *perfect prognosis* assumptions, which are particularly relevant in the case of “black-box” models such as CNNs, whose results cannot be easily explained based on physical reasons, and could potentially lead to implausible downscaled projections due to uncontrolled extrapolation artifacts. As a first step to test the suitability of these models we test potential extrapolation problems by comparing the downscaled and raw model signals. Here we explore this open problem focusing on both temperature and precipitation over Europe and compare the downscaled signals provided by CNNs with those obtained from a set of standard, benchmarking SDMs (in particular different variants of GLMs). Our results show that, as compared to GLMs, CNNs provide more spatially homogeneous downscaled patterns which exhibit smaller departures from the raw GCM outputs for the end of century.

The paper is organized as follows: In Sect. 2 we describe the data, SDMs and indicators used. The results obtained are discussed throughout Sect. 3, with a special focus on the analysis of the climate change signals projected for the end of the century (Sect. 3.3). Finally, the main conclusions are presented in Sect. 4.

2 Data and methods

2.1 Data

We use the reference datasets and periods of analysis proposed in VALUE Experiment 2—an extension of Experiment 1 used in Baño-Medina et al (2019),—which defines a comprehensive framework in the context of climate change. In particular, we use as predictors the variables shown in Table 1 to downscale daily surface temperature and precipitation over Europe, using E-OBS (version 14, at a 0.5° spatial resolution, Cornes et al (2018)) as observational dataset.

For the training phase, predictor data covering the domain 36°N–72°N, 10°W–32°E is taken from the ERA-Interim reanalysis (Dee et al 2011) at a 2° horizontal resolution for the period 1979–2008. For the downscaling phase, we use predictors from the EC-Earth model (*r12i1p1* run)—belonging to the Coupled Models Intercomparison Project (Phase 5) CMIP5 (Hazeleger et al 2010)—for 1979–2008 (historical scenario) and 2071–2100 (RCP8.5 scenario). EC-Earth was re-gridded from its native spatial resolution (1.12°) to the ERA-Interim’s grid (2°) using bilinear interpolation.

Both ERA-Interim and EC-Earth predictor data are available from the VALUE website (<http://www.value-cost.eu/validation>) and can be downloaded as netCDF files. E-OBS data is available at the ECA&D webpage (<https://www.ecad.eu/download/ensembles/download.php>).

Table 1 List of predictor and predictand variables proposed in VALUE, and used in this study

Variable	Units	Levels (hPa)	Predictor	Predictand
Geopotential	m	500, 700, 850 and 1000	✓	
Air temperature	°C	500, 700, 850 and 1000	✓	
Specific humidity	kg/kg	500, 700, 850 and 1000	✓	
Zonal wind velocity	m/s	500, 700, 850 and 1000	✓	
Meridional wind velocity	m/s	500, 700, 850 and 1000	✓	
Surface temperature	°C	–		✓
Total rainfall	mm	–		✓

2.2 Statistical downscaling methods

We selected for this work the CNN1 and CNN10 models presented in Baño-Medina et al (2019) to downscale daily precipitation and temperature, respectively. These models consist of an input layer (with stacked spatial predictors) feeding three layers of convolutions (50:25:1 for precipitation and 50:25:10 for temperature), each formed by 3×3 spatial kernels. The last convolution is fully-connected to the output layer (the E-OBS land-gridpoints), using linear transformations, or sigmoidal ones for probability of rain (i.e. the parameter p of the Bernoulli distributions). The networks are trained to learn daily conditional distributions of precipitation (maximizing the loglikelihood of a mixed Bernoulli-Gamma distribution) and temperature (minimizing the mean square error, MSE), given the predictors; i.e., the network is forced to estimate the associated parameters to the mentioned distributions. For the case of precipitation, we follow a 3-parameter estimation approach which was introduced by Williams (1998) for feedforward neural networks, later adopted by Cannon (2008) and recently extended to CNNs in Baño-Medina et al (2019). In this approach, given a set of large-scale atmospheric predictors, the probability of rain, p , and the shape and scale parameters of a Gamma distribution (α and β , respectively) are simultaneously estimated at each site by minimizing the negative log-likelihood of a Bernoulli-Gamma distribution. For temperature, we minimize the MSE, which is equivalent to minimize the negative log-likelihood of a Gaussian distribution for the conditional mean.

The potential of CNN topologies resides in the efficient treatment of complex spatial features. In climate downscaling, these models have the capacity to handle high-dimensional predictor spaces, automatically selecting the most relevant variables and geographical domains affecting each particular site (see Baño-Medina et al 2019; Baño-Medina 2020, for more details on CNN model interpretability). This aspect is crucial since state-of-the-art SD techniques such as GLMs are unable to treat this high-dimensionality without leading to overfitting, requiring thus some kind of human-guided feature selection (with the consequent loss of relevant information) in most of cases. In fact, to some extent, CNNs

can be viewed as a natural extension to GLMs entailing an intelligent feature selection which allows to maximize the predictive capacity of the available predictor fields.

For these reasons, we only consider as benchmark three GLMs which have been previously used in Bedia et al (2018) and rigorously intercompared in VALUE experiment 1 (Gutiérrez et al 2019). As noted, regression-based SDMs typically undergo feature selection/extraction techniques to select a reduced set of optimal predictors. The GLMs used in this work differ only in the spatial character of the predictors considered. On the one hand, GLM1 (GLM4) uses as predictor local information at the closest (four closest) gridpoint(s) to the site's location. On the other hand, GLMPC builds on the leading principal components (PCs, (Preisendorfer 1988)) explaining the 95% of the variance of the predictor space. The PCs are calculated over the eight PRUDENCE regions as described in Gutiérrez et al (2019). Based on these three predictor configurations, we build three different GLM models at each location; a logistic regression to estimate precipitation occurrence, a Gamma-like regression with logarithmic link to estimate rainfall amount and a Gaussian-like regression to estimate surface temperature.

Though regression-based methods have proved successful to provide unbiased estimates of the mean, they are known to underestimate the local variability, especially for precipitation (Pryor and Schoof 2020). Several approaches have been proposed to alleviate this issue, from variance inflation (Von Storch 1999) to the inference of daily probability distributions conditioned to the given large-scale predictors (see, e.g., Williams 1998; Cannon 2008; Baño-Medina et al 2019). Despite challenges still exist (e.g., preservation of temporal autocorrelation and/or the spatial fields), these approaches have been shown to provide overall good results in terms of local variability. For this reason, we design all our models, both CNNs and GLMs, such that they estimate conditional gaussian (for temperature) and Bernoulli-Gamma (for precipitation) daily probability distributions. As a consequence, both deterministic and stochastic downscaled series can be obtained for all the SDMs developed in this study, depending on whether we limit the prediction to the estimated mean or we perform a sampling from the given daily PDF, respectively. In the case of deterministic precipitation, the binary

series (0/1: no rain/rain) is obtained according to the threshold that matches the rainfall frequency in the train period to the observed value, whereas rainfall amount is directly given as the Gamma's expectation. For both deterministic and stochastic versions, the final predicted precipitation is obtained by multiplying the binary occurrence series by the continuous amount series, the latter according to the conditional Gamma PDFs. For the case of temperature, only deterministic implementations of both CNNs and GLMs are considered, since the local variability of this variable is lower and stochasticity is not really needed.

Table 2 summarizes the dimensionality of the models used in this work. Whereas CNN1 and CNN10 models are designed over a predictor space of 19×22 (latitude-longitude) with the 20 variables listed in Table 1 stacked as inputs (or channels, similar to RGB channels in computer vision images), GLMs take as inputs the predictor variables at the closest gridpoints to each site (20 and 80 for GLM1 and GLM4, respectively) or the PCs that explain the 95% of the total variance of the predictor space (this depends on the PRUDENCE region so we marked it with an 'x' in Table 2). In CNN1/10, the input layer is convolutionally connected to a set of 3 hidden layers with 50, 25 and 1/10 feature maps in a sequential manner. Padding is applied in the CNN1 model, keeping the input resolution constant, whilst CNN10 lack of this property, diminishing the spatial resolution throughout the hidden layers as a function of the kernel size (3×3). The output layer of the CNNs matches the E-OBS land-gridpoints resulting into 3259 predictand sites for the precipitation/temperature model. Due to the nature of the GLM optimization and since the predictor spaces do not overlap for the sites of interest, GLMs are formulated in single-site mode—one model per site,—in contrast with CNNs that operate in multi-site mode—all sites are simultaneously

downscaled from a single precipitation (CNN1) or temperature (CNN10) model. The capacity of DL topologies to downscale to multiple sites at a time is commonly referred to as multi-task learning, and its benefits include: computational efficiency, implicit regularization and ability to estimate multivariate distributions (see Ruder 2017, for a review in multi-task deep learning). On the one hand, computational efficiency comes from the fact that a single model is trained and used to predict over the test samples. On the other hand, multi-task topologies forces the network to learn patterns that are useful to downscale at multiple sites, sharing knowledge in the hidden layers, and acting as a form of inductive bias that prevents the network from overfitting. In particular, this regularization property was tested for the topologies used in this study in Baño-Medina and Gutiérrez (2019), where multi-site CNNs attained better results than their equivalent single-site CNNs in terms of local reproducibility. However, it has to be noted that none of these benefits imply an improvement on the spatial consistency of the downscaled fields, and explicit modeling of multivariate distributions—e.g., daily multivariate Gaussian distributions—in addition to multi-site networks is needed (Cannon 2008).

2.3 Validation indices

We use in this work some of the indicators that have been consolidated in the VALUE validation framework (Maraun et al 2015). In particular, we consider P02, Mean and P98 for temperature and R01, SDII and P98 for precipitation (see Table 3 for details).

In particular, to assess the performance of the downscaled results obtained from EC-Earth in its historical scenario (1979–2008), we compute the biases (with respect to

Table 2 Dimensions of the input (In), output (Out) and hidden (H1, H2, and H3) layers for the methods considered in this study applied at a continental scale over Europe

	In	H1	H2	H3	Out	Models
GLM1	20	-	-	-	1	3259
GLM4	80	-	-	-	1	3259
GLMPC	x	-	-	-	1	3259
CNN1	(19, 22, 20)	(19, 22, 50)	(19, 22, 25)	(19, 22, 1)	3259	1
CNN10	(19, 22, 20)	(17, 20, 50)	(15, 18, 25)	(13, 16, 10)	3259	1

Table 3 List of indicators used in this study

Code	Description	Units	Variable
R01	Frequency of wet (precip. ≥ 1 mm/day) days	%	P
SDII	Simple daily intensity index (mean precip. in wet days)	mm/day	P
P02	2nd percentile	°C	T
Mean	Mean	°C	T
P98	98th percentile (for precip., only wet days are considered)	°C, mm/day	T, P

“P” and “T” denote precipitation and temperature, respectively

E-OBS) for the mentioned indicators. For temperature (precipitation), absolute (relative, in %) biases are considered. Differently, for the case of the downscaled results obtained from EC-Earth in the far future, we compute the delta change between the RCP (2071-2100) and the historical (1979-2008) scenario—for each temperature (precipitation) indicator, absolute (relative, in %) differences are given. In this regard, note that, although downscaling may be expected to modify the raw GCM projections at the local scale (due, e.g., to a better modeling of local phenomena such as orography), neither the spatial structure nor the magnitude of the raw climate change signal should be significantly altered over a sufficiently large region (see Manzanas et al 2020, for details). With this in mind, we also compute the existing differences in the delta changes as projected by the downscaled version and by the raw GCM ($\Delta_{down} - \Delta_{GCM_{raw}}$). As such, for a particular SDM, large differences would indicate a bad extrapolation capability—understood as compatibility with the raw GCM projections.

3 Results

In this section we discuss the suitability of the different statistical methods presented in Sect. 2.2 (with a special interest in the CNNs) to generate plausible regional climate change scenarios.

3.1 Testing the perfect-prognosis assumption

We first assess if the predictors used in our SDMs are realistically simulated by the EC-Earth. Note that this model has been shown to consistently reproduce key large scale processes affecting the European climate, in particular storm tracks (Lee 2015), which makes it a suitable option for downscaling, as proposed in VALUE.

To avoid the potential issues related to the presence of systematic biases in the first and second moments of the GCM (see, e.g., Vrac and Vaittinada Ayar 2016; Nikulin et al 2018; Manzanas et al 2019), we used standardized anomalies (at a gridbox level) for both EC-Earth and ERA-Interim. Moreover, in order to avoid also the possible misrepresentation of the annual cycle, we have also assessed the effect of applying a simple monthly mean bias adjustment (BA) prior to standardization. BA consists in adjusting the EC-Earth (x_{GCM}) monthly means towards the corresponding reanalysis values (x_{REA}), gridbox by gridbox (Eqs. 1, where j refers to a particular variable, $i = 1, 2, \dots, 12$ to the month of the year, and h and f denote historical and RCP periods, respectively). The reader is referred to Gutiérrez et al (2019) for further details.

$$x_{j,GCM_h}^i = x_{j,GCM_h}^i - \bar{x}_{GCM_h}^i + \bar{x}_{REA}^i \quad (1)$$

$$x_{j,GCM_f}^i = x_{j,GCM_f}^i - \bar{x}_{GCM_h}^i + \bar{x}_{REA}^i \quad (2)$$

Figure 1 allows to assess the distributional similarity between the ERA-Interim's and EC-Earth's temporal series over the historical period 1979–2008. In particular, it shows the results from a Kolmogorov–Smirnov test for standardized (left) and bias adjusted+standardized (right) daily predictors over the entire year, winter and summer (in columns). Red crosses identify those gridpoints where the null hypothesis of the test—ERA-Interim and EC-Earth distributions are indistinguishable—can be rejected at a 5% significance level. In all cases, colors show the p -values (in the range 0–0.3) corresponding to the ERA-Interim vs. raw (with no transformation) EC-Earth comparison. For brevity, results are only shown for two illustrative variables, temperature at 1000 hPa (T1000) and specific humidity at 700 hPa (Q700), top and bottom row, respectively.

Both T1000 and Q700 present in general low p -values (below the significance level of 0.05), reflecting that EC-Earth and ERA-Interim distributions are significantly different over many regions. This is mainly due to the presence of systematic biases in EC-Earth, since the situation is substantially improved once standardization is carried out, regardless BA is applied or not (see the red crosses in both panels). If reanalysis and GCM predictors are compared over the entire year (annual distributions, left column in each panel) there is in general good distributional agreement for T1000 and Q700 over the domain (with a few exceptions in the Mediterranean for the case of Q700). However, when the comparison is undertaken for winter and summer (middle and right column in each panel, respectively), better results are found when BA is applied. These results prove that monthly bias adjustment helps to meet the *perfect prognosis* assumption, yielding better predictors for downscaling. Moreover, though not showed here for brevity, for other predictor variables—especially wind velocity components in southern Europe and specific humidity at other height levels,—BA is crucial to make reanalysis and GCM predictors compatible.

3.2 Downscaling performance in the historical period

As explained in the previous section, the SDMs introduced in Sect. 2.2 are first trained using ERA-Interim standardized predictors and subsequently applied to EC-Earth predictors, after bias adjustment and standardization. Figure 2 shows the validation results obtained for the indices listed in Table 3 for the historical period 1979–2008, calculated as relative (absolute) biases for precipitation (temperature). Results

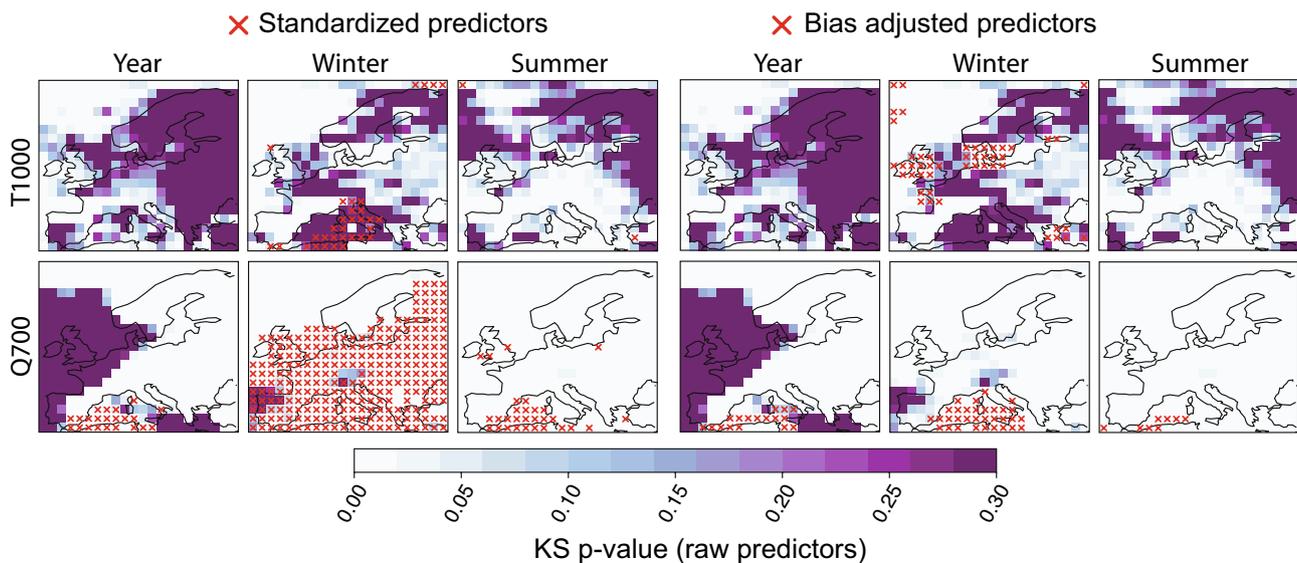


Fig. 1 Results from a Kolmogorov-Smirnov (KS) test for standardized (left) and bias adjusted+standardized (right) daily T1000 and Q700 (top and bottom row, respectively) over the entire year, winter and summer (in columns). Red crosses identify those gridpoints where the null hypothesis of the test—ERA-Interim and EC-Earth

distributions are indistinguishable—can be rejected at a 5% significance level. In all cases, colors show the p-values (in the range 0 – 0.3) corresponding to the ERA-Interim vs. raw (with no transformation) EC-Earth comparison

are shown for the raw EC-Earth outputs (first row) and for the different statistical downscaling methods (GLMs: rows 2–4 and CNN: row 5), considering E-OBS as the observational reference in all cases. This figure shows that EC-Earth exhibits moderate to large biases for both precipitation and temperature over vast parts of Europe, with a tendency to overestimate precipitation occurrence (the well-known ‘drizzle effect’) and underestimate precipitation intensity and extremes (indicating a systematic shrinkage of the distribution). For temperature, EC-Earth underestimates the mean and extremes in the Mediterranean (indicating a systematic shift of the distribution) and under/over-estimates the warm/cold extremes in regions of central and Northern Europe (indicating a systematic shrinkage).

All the SDMs considered, largely reduce the biases encountered for centered statistics such as R01 (for precipitation) and the mean (for temperature). This is not surprising as they are designed to minimize the mean errors (w.r.t. the E-OBS observations) during the training process. An exception to this is precipitation intensity for those methods relying on local predictors, which overestimate intensity (particularly GLM4). Nevertheless, this problem is alleviated when spatial predictors are used, either PCs in the case of GLMPC or convolutions in the CNN.

In the case of precipitation, all SDMs underestimate extreme values (P98), which is due to the reduced local variability explained by large-scale predictors—smaller underestimation corresponds to those methods with a presence of more informative variables in the predictor set i.e., GLM4

and CNN (Baño-Medina et al 2019). Figure 2 shows two columns for P98, corresponding to the deterministic (DET.) and stochastic (STOCHASTIC) versions described in Sect. 2.2. With the exception of the GLMs using local predictors (in particular GLM4), a clear improvement is found for the stochastic versions, which yield substantially lower biases.

For temperature, all SDMs yield nearly negligible biases in most cases (especially for the mean) and extreme warm temperatures (P98). However, the three GLM-based implementations overestimate P2 in Scandinavia, where the EC-Earth model exhibits the largest biases. To a great extent, this is corrected by the CNN, which points out again the benefit of convolutional networks for climate downscaling purposes.

In agreement with the results previously found in Baño-Medina et al (2019), we confirm here that CNNs provide overall better results than GLMs. This is due to the inherent capacity of CNNs to automatically extract the important spatial features determining the local climate, which allows to properly model the complex relationships (both in space and in time) that are established between the local- and the large-scale and improves the out-of-sample generalization capacity, especially for precipitation.

3.3 Future climate projections: raw and downscaled signals

A key assumption for the secure application of statistical downscaling to produce climate change projections is that

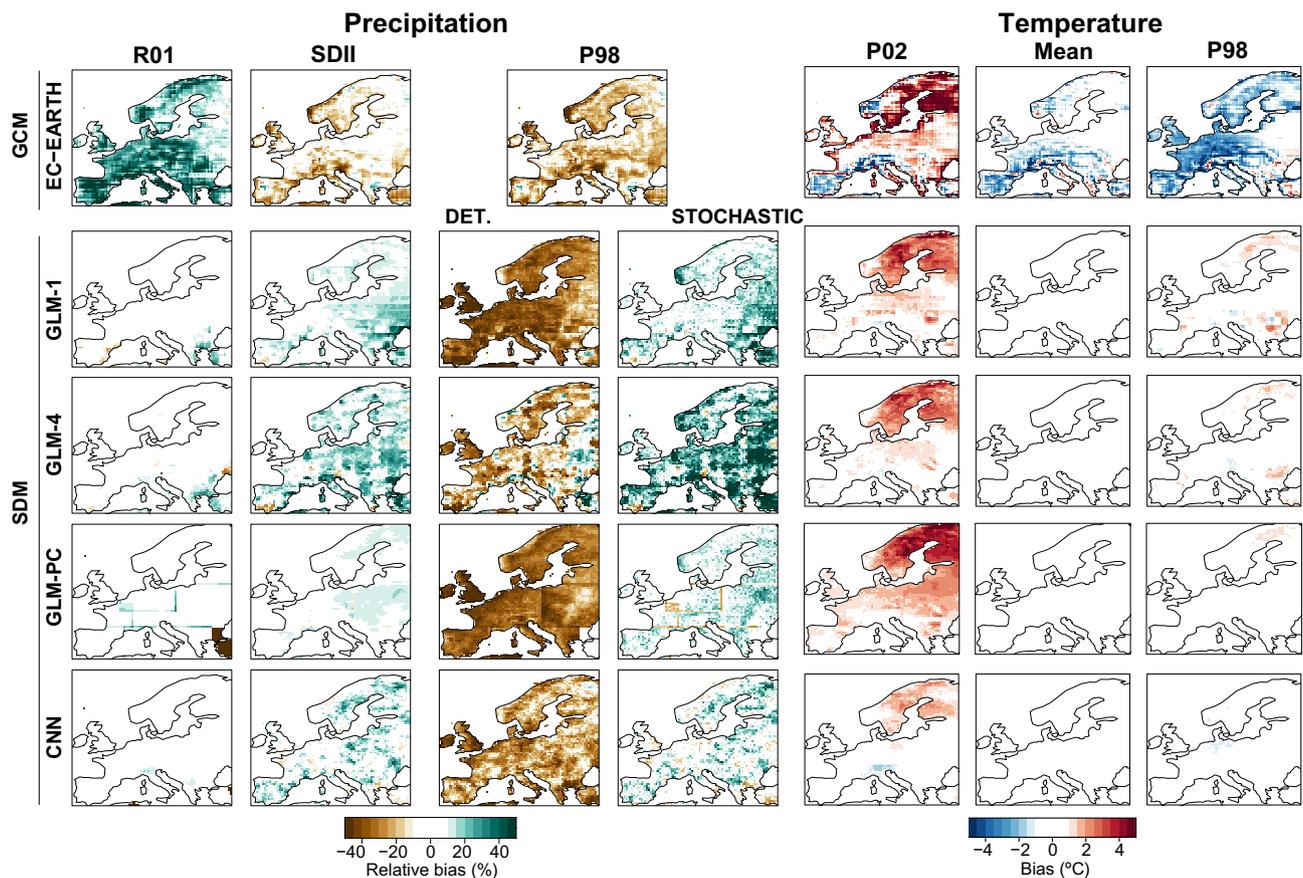


Fig. 2 Biases for the indicators related to precipitation (left) and temperature (right) listed in Table 3, as obtained from the raw EC-Earth simulations (row 1) and the different SDMs considered (GLMs: rows 2–4 and CNN: row 5) for the historical period 1979–2008. In all cases, the observational reference used is E-OBS

SDMs should be able to generalize and extrapolate to previously unseen (e.g. climate change) conditions—stationarity assumption. Following the recommendations done by Gutiérrez et al (2013), this was partially analyzed in Baño-Medina et al (2019) using an anomalous warm test period in comparison to that observed during training, obtaining consistent and unbiased downscaled predictions. However, a more robust analysis is needed to assess potential problems that may arise in the downscaled climate change signal produced by the different SDMs (as compared to the one given by the raw GCMs).

The first row in Fig. 3 shows the delta changes projected by the EC-Earth for 2071–2100 (with respect to the baseline period 1979–2008), considering its raw outputs under the RCP8.5 scenario. For the precipitation (temperature) indicators, shown in the left (right) panel, relative (absolute) values are displayed.

According to this GCM, a decrease in rainfall frequency (R01) might be expected over the Mediterranean, whereas the intensity (SDII) would increase in mid and northern Europe. Extreme precipitation (as represented by P98)

would increase all over the area of study. Temperature is projected to rise significantly all over Europe, up to 5°C for the mean, but reaching even higher increases for extreme temperatures (P2 and P98) in northern and southern Europe, respectively. We want to remark that the goal here is not providing comprehensive climate change scenarios over Europe (which should build on multi-model ensembles), but to assess the impact that different techniques may have on the downscaled projections.

Note that, unless it can be justified by process understanding, significant deviations from the the global model’s climate change signal over large regions could be an indicator of physically inconsistent and implausible downscaled results (Manzanas et al 2020). With this in mind, rows two to five in Fig. 3 show the differences between the downscaled (not shown) and EC-Earth (first row) delta changes, as given by the different SDMs considered (GLM: rows 2–4, CNN: row 5). Absolute (relative, in %) differences are shown for the case of temperature (precipitation). White colors represent regions where the SDMs preserve the climate change signal given by EC-Earth, whereas brown/blue or green/red

would increase all over the area of study. Temperature is projected to rise significantly all over Europe, up to 5°C for the mean, but reaching even higher increases for extreme temperatures (P2 and P98) in northern and southern Europe, respectively. We want to remark that the goal here is not providing comprehensive climate change scenarios over Europe (which should build on multi-model ensembles), but to assess the impact that different techniques may have on the downscaled projections.

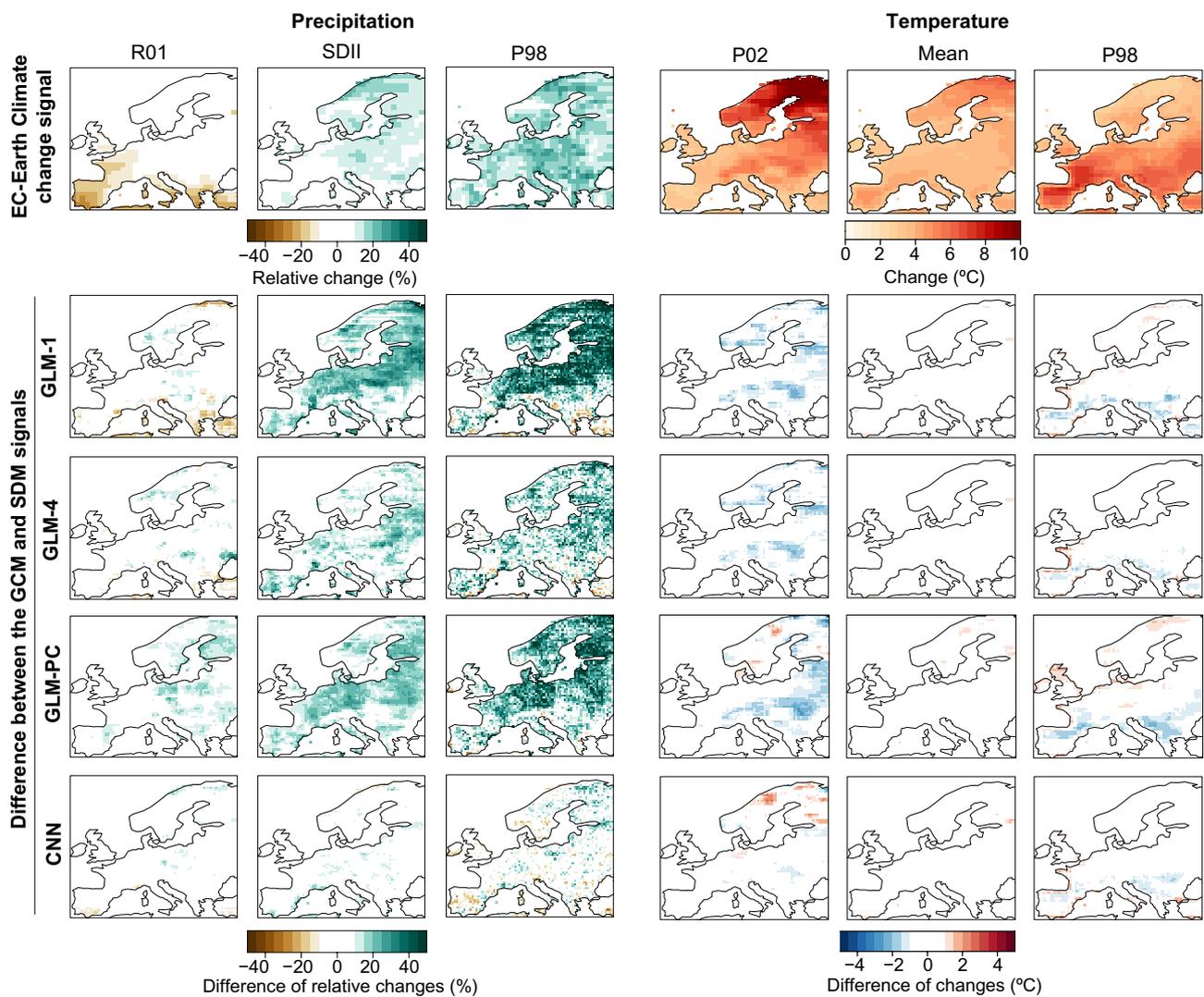


Fig. 3 The first row shows the delta changes projected by the EC-Earth for 2071–2100 (with respect to the baseline period 1979–2008), considering its raw outputs under the RCP8.5 scenario. Rows two to five in Fig. 3 show the differences between the downscaled (not

shown) and EC-Earth (first row) delta changes, as given by the different SDMs considered (GLM: rows 2–4, CNN: row 5). For the precipitation (temperature) indicators, shown in the left (right) panel, relative (absolute) values are displayed

colors indicate regions where downscaling reduce or enlarge it, respectively.

For precipitation, both GLMs and CNN preserve the climate change signal for frequency (R01). However, GLMs (particularly GLM1 and GLMPC) enlarge the mean and extreme precipitation signals up to 20% and 40% for SDII and P98, respectively. In principle, there is no known physical mechanisms supporting such changes and could, therefore, be attributed to the overestimation of both indicators found for GLMs in Sect. 3.2. Nevertheless, the pattern found for biases in the historical period (Fig. 2) does not match with the climate change signal shown in Fig. 3. Differently, the CNN largely preserves the climate change signal given by the global model for all the indices considered, posing

no challenges on the interpretation of the patterns obtained. A similar situation is found for temperature, for which CNN preserves to a great extent the global model’s climate change signal both for the mean and extremes, with GLMs exhibiting some reduction in small regions over central and northern Europe (especially for P2).

In contrast to what happens for standard GLMs, the results from this work evidence that CNNs provide plausible downscaled information (when fed by GCM predictors) for the provision of regional-to-local climate change scenarios.

4 Conclusions

Recently, Baño-Medina et al (2019) assessed the performance of convolutional neural networks (CNNs) for perfect prog statistical downscaling over Europe using “perfect” reanalysis data as predictors. Their results showed that CNNs can efficiently work with continental-sized domains, outperforming other well-established statistical models for particular forecast aspects. We extend here this work by analyzing the suitability of CNNs for downscaling climate change, applying the models to predictors from future GCM climate projections.

As a first step we assess the performance of CNNs to downscale temperature and precipitation from the historical scenario of the EC-Earth model. For completeness, we also include in the analysis three different implementations of generalized linear models (GLMs), which ranked amongst the best ones in the VALUE intercomparison experiment (Gutiérrez et al 2019). Our results indicate that statistical downscaling (and in particular CNNs) allows to reduce the systematic errors that are usually present in GCMs for the mean and extremes, providing more realistic climate information. We found that, as compared to methods based on spatial predictors, GLMs based on local predictors are more sensitive to the possible inconsistencies that may arise between reanalysis and GCM predictor data, yielding higher biases, particularly for precipitation amount metrics i.e., SDII and P98. This seems reasonable since these inconsistencies among datasets are directly fed to the models if no manipulation of the predictor space is carried out in the form of e.g., convolutional layers, resulting in deficiencies of the “perfect-prognosis” condition.

In a second step, we study whether or not CNNs provide a suitable alternative for the generation of reliable local to regional downscaled climate change scenarios, which, to the author’s knowledge, has not yet been explored. We compare the downscaled climate change signals produced by CNNs with those obtained from the benchmarking GLMs. The suitability of the different methods tested for climate change applications is quantified based on the similarity with the raw projections given by the EC-Earth (under the RCP8.5 scenario). GLMs are found to yield local scenarios which are not fully consistent with the signals produced by the GCM, especially for the case of precipitation. Differently, the projections given by CNNs are comparable (to a great extent) to the change signals provided by EC-Earth. This suggests the adequacy of CNNs for the downscaling of local-to-regional climate change scenarios building on the good generalization properties and stable behaviors under climate change conditions.

The results from this work may foster the use of CNNs for the generation of reliable climate change information

on continental-sized domains, which is crucial for the implementation of adequate mitigation policies. To further corroborate these conclusions, we plan to extend the present work to other geographical regions and variables with different climatological properties under the umbrella of the international initiative CORDEX-ESD, whose objective is to produce high-resolution climate change information worldwide.

Appendix: Reproducibility of results

All the data used in this work (E-OBS observations, ERA-Interim and EC-Earth projections) are publicly available and accessible from the User Data Gateway (UDG), a THREDDS-based service from the Santander Climate Data Service which provides access to a wide catalog of popular climate datasets. These datasets can be remotely accessed using the open *climate4R* framework (Iturbide et al 2019). See also <https://github.com/SantanderMetGroup/climate4R> for a complete description of the different packages forming this framework.

The standard SDMs considered here (different implementations of GLMs) are built with the *downscaleR* package (Bedia et al 2019), and the convolutional deep models are built using *downscaleR.keras* (<https://github.com/SantanderMetGroup/downscaleR.keras>), a wrapper that integrates *keras*—the state-of-the-art library in deep learning— within *climate4R*. To validate the predictions/projections, we use the set of indices defined in VALUE (see <http://www.value-cost.eu> for details), which are available in the *climate4R.value* package.

The companion Jupyter notebook, accessible from the *deepDownscaling* GitHub repository of the Santander Meteorology Group (<https://github.com/SantanderMetGroup/DeepDownscaling>), describes all the steps necessary to fully reproduce the results presented in this manuscript, which were produced on a machine running under Ubuntu 18.04.3 LTS (64 bits), with 60 GiB memory and a multi-core CPU composed of 16 processing units and 32 threads Intel(R) Xeon(R) CPU E5-2670 of 2.60 GHz. The computational times needed to train and predict in both historical and RCP8.5 scenarios are described in Table 4 for the statistical models tested. Note that these times do not entirely

Table 4 Computational times needed to train and predict on both historical and RCP8.5 scenarios by the GLM and CNN models (in columns) for the case of precipitation (P) and temperature (T)

	GLM1	GLM4	GLMPC	CNN
P	4h 15min	8h 21min	2h 30min	1h 11min
T	2h 32min	2h 20min	1h 20min	50min

depend on the method's nature—e.g., GLMs are solved analytically while CNNs required an iterative optimization procedure,—but also on the internal R libraries used to build the models, and on its single- or multi-site nature—e.g., only one model is trained for the CNNs.—Overall we observe that GLMs consume more resources and timing than CNNs, in particular for precipitation downscaling. Whilst CNNs estimate simultaneously the occurrence and quantity of rain through the Bernoulli–Gamma conditional PDFs, two independent models are needed for the GLMs, what increases considerably the computational times. This is ratified by comparing the results obtained for precipitation and temperature GLMs; the latter requiring half of the time used by the former (although still overpassing the CNN). In addition, since for the GLMs each site is downscaled with different predictor spaces, all data processing, training and prediction steps need to be done by chunks of latitudes, leading thus to notably longer times than CNNs.

Acknowledgements We acknowledge the E-OBS dataset from the EU-FP6 project UERRA (<http://www.uerra.eu>) and the Copernicus Climate Change Service, and the data providers in the ECA&D project (<https://www.ecad.eu>). The authors acknowledge partial support from the ATLAS project from the Spanish Research Program (AEI; PID2019-111481RB-I00).

Funding The authors acknowledge partial support from the ATLAS project, funded by the Spanish Research Program (PID2019-111481RB-I00). Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Availability of data and material On the one hand, all the data involved in the Experiment 2 of the COST action VALUE—in particular the ERA-Interim and EC-Earth variables used in this work—are available at <http://www.value-cost.eu/data#netcdf>. On the other hand, the E-OBS dataset can be downloaded from the ECA&D webpage: <https://www.ecad.eu/download/ensembles/download.php>.

Declarations

Conflicts of interest Not applicable.

Code availability For the purpose of research transparency, we provide a Jupyter notebook which allows to fully reproduce the results presented herein. It can be reached at the Santander Meteorology Group GitHub repository (https://github.com/SantanderMetGroup/DeepDownscaling/blob/master/2020_Bano_CD.ipynb, last access: 05/03/2021, DOI: 10.5281/zenodo.4580590).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will

need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Baño-Medina J (2020) Understanding Deep Learning Decisions in Statistical Downscaling Models, Association for Computing Machinery, New York, NY, USA, p 79–85. <https://doi.org/10.1145/3429309.3429321>
- Baño-Medina J, Gutiérrez JM (2019) The importance of inductive bias in convolutional models for statistical downscaling. In: Proceedings of the 9th international workshop on climate informatics: CI 2019, <https://doi.org/10.5065/y82j-f154>, https://github.com/SantanderMetGroup/DeepDownscaling/blob/master/2019_Bano_CI.pdf
- Baño-Medina J, Manzanos R, Gutiérrez JM (2019) Configuration and Intercomparison of deep learning neural models for statistical downscaling. preprint, 10.5194/gmd-2019-278
- Bedia J, Iturbide M, Herrera García S, Baño-Medina J, Fernández J, Frías M, Manzanos R, San-Martín D, Cimadevilla E, Cofiño A, Gutiérrez J (2018) The R-based climate4R open framework for reproducible climate data access and post-processing. Environmental Modelling and Software. <https://doi.org/10.1016/j.envsoft.2018.09.009>
- Bedia J, Baño-Medina J, Legasa MN, Iturbide M, Manzanos R, Herrera S, Casanueva A, San-Martín D, Cofiño AS, Gutiérrez JM (2019) Statistical downscaling with the downscaleR package: Contribution to the VALUE intercomparison experiment. preprint, Climate and Earth System Modeling, 10.5194/gmd-2019-224, <https://www.geosci-model-dev-discuss.net/gmd-2019-224/>
- Cannon AJ (2008) Probabilistic multisite precipitation downscaling by an expanded Bernoulli–Gamma density network. J Hydro-meteorol 9(6):1284–1300
- Cornes RC, van der Schrier G, van den Besselaar EJM, Jones PD (2018) An ensemble version of the E-OBS temperature and precipitation data sets. J Geophys Res 123(17):9391–9409
- Dee DP, Uppala SM, Simmons AJ, Berrisford P, Poli P, Kobayashi S, Andrae U, Balmaseda MA, Balsamo G, Bauer P, Bechtold P, Beljaars ACM, Lvd Berg, Bidlot J, Bormann N, Delsol C, Dragani R, Fuentes M, Geer AJ, Haimberger L, Healy SB, Hersbach H, H \ddot{A} ³lm EV, Isaksen I, K \ddot{A} llberg P, K \ddot{A} hler M, Matricardi M, McNally AP, Mongeá Sanz BM, Morcrette JJ, Park BK, Peubey C, Rosnay Pd, Tavolato C, Thépaut JN, Vitart F, (2011) The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Quarterly Journal of the Royal Meteorological Society 137(656):553–597
- Gutiérrez JM, San-Martín D, Brands S, Manzanos R, Herrera S (2013) Reassessing statistical downscaling techniques for their robust application under climate change conditions. J Clim 26(1):171–188
- Gutiérrez JM, Maraun D, Widmann M, Huth R, Hertig E, Benestad R, Roessler O, Wibig J, Wilcke R, Kotlarski S, Martín DS, Herrera S, Bedia J, Casanueva A, Manzanos R, Iturbide M, Vrac M, Dubrovsky M, Ribalaygua J, Pórtoles J, R \ddot{a} ty O, R \ddot{a} isänen J, Hingray B, Raynaud D, Casado MJ, Ramos P, Zerenner T, Turco M, Bosshard T, Ātp \acute{a} nek P, Bartholy J, Pongracz R, Keller DE, Fischer AM, Cardoso RM, Soares PMM, Czernecki B, Pagé C (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>

- Hazeleger W, Severijns C, Semmler T, Å tefÄ nescu S, Yang S, Wang X, Wyser K, Dutra E, Baldasano JM, Bintanja R, Bougeault P, Caballero R, Ekman AML, Christensen JH, van den Hurk B, Jimenez P, Jones C, Källberg P, Koenigk T, McGrath R, Miranda P, van Noije T, Palmer T, Parodi JA, Schmith T, Selten F, Storelvmo T, Sterl A, Tapamo H, Vancoppenolle M, Viterbo P, Willén U, (2010) EC-Earth: A Seamless Earth-System Prediction Approach in Action. *Bulletin of the American Meteorological Society* 91(10):1357–1364. <https://doi.org/10.1175/2010BAMS2877.1>
- Iturbide M, Bedia J, Herrera S, Baño-Medina J, Fernández J, Frías MD, Manzananas R, San-Martín D, Cimadevilla E, Cofiño AS, Gutiérrez JM, (2019) The R-based climate4R open framework for reproducible climate data access and post-processing. *Environ Model Softw* 111:42–54
- Jacob D, Teichmann C, Sobolowski S, Katragkou E, Anders I, Belda M, Benestad R, Boberg F, Buonomo E, Cardoso RM, Casanueva A, Christensen OB, Christensen JH, Coppola E, De Cruz L, Davin EL, Dobler A, Domínguez M, Fealy R, Fernandez J, Gaertner MA, García-Díez M, Giorgi F, Gobiet A, Goergen K, Gómez-Navarro JJ, Alemán JGG, Gutiérrez C, Gutiérrez JM, Güttler I, Haensler A, Halenka T, Jerez S, Jiménez-Guerrero P, Jones RG, Keuler K, Kjellström E, Knist S, Kotlarski S, Maraun D, van Meijgaard E, Mercogliano P, Montávez JP, Navarra A, Nikulin G, de Noblet-Ducoudré N, Panitz HJ, Pfeifer S, Piazza M, Pichelli E, Pietikäinen JP, Prein AF, Preuschmann S, Rechid D, Rockel B, Romera R, Sánchez E, Sieck K, Soares PMM, Somot S, Srnec L, Sørland SL, Termonia P, Truhetz H, Vautard R, Warrach-Sagi K, Wulfmeyer V (2020) Regional climate downscaling over Europe: perspectives from the EURO-CORDEX community. *Regional Environ Change* 20(2):51. <https://doi.org/10.1007/s10113-020-01606-9>
- Lee RW (2015) Storm track biases and changes in a warming climate from an extratropical cyclone perspective using cmip5, <http://centaur.reading.ac.uk/79416/>
- Manzananas R, Brands S, San-Martín D, Lucero A, Limbo C, Gutiérrez JM (2015) Statistical downscaling in the tropics can be sensitive to reanalysis choice: a case study for precipitation in the philippines. *J Clim* 28(10):4171–4184. <https://doi.org/10.1175/JCLI-D-14-00331.1>
- Manzananas R, Gutiérrez JM, Bhend J, Hemri S, Doblas-Reyes FJ, Torralba V, Penabad E, Brookshaw A (2019) Bias adjustment and ensemble recalibration methods for seasonal forecasting: a comprehensive intercomparison using the C3S dataset. *Clim Dyn* 53(3):1287–1305. <https://doi.org/10.1007/s00382-019-04640-4>
- Manzananas R, Fiwa L, Vanya C, Kanamaru H, Gutiérrez JM (2020) Statistical downscaling or bias adjustment: a case study involving implausible climate change projections of precipitation in Malawi. *Clim Change* 162(3):1437–1453. <https://doi.org/10.1007/s10584-020-02867-3>
- Maraun D, Widmann M (2018) *Statistical Downscaling and Bias Correction for Climate Research*. Cambridge University Press, google-Books-ID: AMhJDwAAQBAJ
- Maraun D, Widmann M, Gutiérrez JM, Kotlarski S, Chandler RE, Hertig E, Wibig J, Huth R, Wilcke RA (2015) VALUE: A framework to validate downscaling approaches for climate change studies. *Earth's Future* 3(1):2014EF000,259, 10.1002/2014EF000259
- Nikulin G, Asharaf S, no MEM, Calmanti S, Cardoso RM, Bhend J, Fernández J, Frías MD, Fröhlich K, Fráh B, Herrera S, Manzananas R, Gutiérrez JM, Hansson U, Kolax M, Liniger MA, Soares PM, Spirig C, Tome R, Wyser K, (2018) Dynamical and statistical downscaling of a global seasonal hindcast in eastern Africa. *Climate Services* 9:72–85
- Pan B, Hsu K, AghaKouchak A, Sorooshian S (2019) Improving precipitation estimation using convolutional neural network. *Water Resources Res* 55(3):2301–2321
- Preisendorfer RW (1988) *Principal component analysis in meteorology and oceanography*, 1st edn. Elsevier, Amsterdam
- Pryor SC, Schoof JT (2020) Differential credibility assessment for statistical downscaling. *J Appl Meteorol Climatol* 59(8):1333–1349
- Ruder S (2017) An overview of multi-task learning in deep neural networks
- San-Martín D, Manzananas R, Brands S, Herrera S, Gutiérrez JM (2016) Reassessing model uncertainty for regional projections of precipitation with an ensemble of statistical downscaling methods. *J Clim* 30(1):203–223
- Sun L, Lan Y (2020) Statistical downscaling of daily temperature and precipitation over china using deep learning neural models: localization and comparison with other methods. *Int J Climatol*. <https://doi.org/10.1002/joc.6769>
- Von Storch H (1999) On the use of ‘inflation’ in statistical downscaling. *Journal of Climate - J CLIMATE* 12:3505–3506. [https://doi.org/10.1175/1520-0442\(1999\)012<3505:OTUOII>2.0.CO;2](https://doi.org/10.1175/1520-0442(1999)012<3505:OTUOII>2.0.CO;2)[https://doi.org/10.1175/1520-0442\(1999\)012<3505:OTUOII>2.0.CO;2](https://doi.org/10.1175/1520-0442(1999)012<3505:OTUOII>2.0.CO;2)
- Vrac M, Vaittinada Ayar P (2016) Influence of bias correcting predictors on statistical downscaling models. *J Appl Meteorol Climatol* 56(1):5–26
- Williams PM (1998) *Modelling Seasonality and Trends in Daily Rainfall Data*. In: Jordan MI, Kearns MJ, Solla SA (eds) *Advances in neural information processing systems* 10, MIT Press, pp 985–991, <http://papers.nips.cc/paper/1429-modelling-seasonality-and-trends-in-daily-rainfall-data.pdf>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.