Geosci. Model Dev., 13, 2109–2124, 2020
https://doi.org/10.5194/gmd-13-2109-2020

Geoscientific Model Development
Open Access
EGU

Configuration and intercomparison of deep learning neural models for statistical downscaling

Jorge Baño-Medina[1], **Rodrigo Manzanas**[2], **and José Manuel Gutiérrez**[1]

[1]Santander Meteorology Group, Institute of Physics of Cantabria, CSIC-University of Cantabria, Santander, Spain
[2]Santander Meteorology Group, Dpto. de Matemática Aplicada y Ciencias de la Computación, Universidad de Cantabria, Santander, Spain

Correspondence: Jorge Baño-Medina (bmedina@ifca.unican.es)

Received: 26 September 2019 – Discussion started: 28 October 2019
Revised: 28 February 2020 – Accepted: 19 March 2020 – Published: 28 April 2020

Abstract. Deep learning techniques (in particular convolutional neural networks, CNNs) have recently emerged as a promising approach for statistical downscaling due to their ability to learn spatial features from huge spatiotemporal datasets. However, existing studies are based on complex models, applied to particular case studies and using simple validation frameworks, which makes a proper assessment of the (possible) added value offered by these techniques difficult. As a result, these models are usually seen as black boxes, generating distrust among the climate community, particularly in climate change applications.

In this paper we undertake a comprehensive assessment of deep learning techniques for continental-scale statistical downscaling, building on the VALUE validation framework. In particular, different CNN models of increasing complexity are applied to downscale temperature and precipitation over Europe, comparing them with a few standard benchmark methods from VALUE (linear and generalized linear models) which have been traditionally used for this purpose. Besides analyzing the adequacy of different components and topologies, we also focus on their extrapolation capability, a critical point for their potential application in climate change studies. To do this, we use a warm test period as a surrogate for possible future climate conditions.

Our results show that, while the added value of CNNs is mostly limited to the reproduction of extremes for temperature, these techniques do outperform the classic ones in the case of precipitation for most aspects considered. This overall good performance, together with the fact that they can be suitably applied to large regions (e.g., continents) without worrying about the spatial features being considered as predictors, can foster the use of statistical approaches in international initiatives such as Coordinated Regional Climate Downscaling Experiment (CORDEX).

1 Introduction

The coarse spatial resolution and systematic biases of global climate models (GCMs) are two major limitations for the direct use of their outputs in many sectoral applications, such as hydrology, agriculture, energy or health, particularly for climate change impact studies (Maraun and Widmann, 2017). These applications typically involve the use of sectoral models (e.g., crop or hydrological models) and/or climate indices (e.g., frost days or warm spells) which require regional to local weather (daily) series of different variables (precipitation, temperature, radiation, wind etc.) over multiple decades representative of the historical and future climates (see, e.g., Galmarini et al., 2019; Ba et al., 2018; Sanderson et al., 2017; Teutschbein et al., 2011; Wang et al., 2017). Moreover, the results of these studies are sensitive to different aspects of the climate data, such as the temporal structure (e.g., in agriculture or energy), the spatial and/or inter-variable structure (e.g., in hydrology) or the extremes (e.g., in hydrology and health).

In order to bridge this gap, different *statistical downscaling* (SD; Maraun and Widmann, 2017) methods have been developed building on empirical relationships established between informative large-scale atmospheric variables (predictors) and local/regional variables of interest (predictands). Under the perfect-prognosis approach, these relationships

are learned from (daily) data using simultaneous observations for both the predictors (from a reanalysis) and predictands (historical local or gridded observations), and are subsequently applied to GCM-simulated predictors (multidecadal climate change projections under different scenarios), to obtain locally downscaled values (see, e.g., Gutiérrez et al., 2013; Manzanas et al., 2018).

A number of standard perfect-prognosis SD (hereafter just SD) techniques have been developed during the last 2 decades building mainly on (generalized) linear regression and analog techniques (Gutiérrez et al., 2018). These standard approaches are widely used by the downscaling community, and several intercomparison studies have been conducted to understand their advantages and limitations, taking into account a number of aspects such as temporal structure, extremes or spatial consistency. In this regard, the VALUE (Maraun et al., 2015) initiative proposed an experimental validation framework for downscaling methods and conducted a comprehensive intercomparison study over Europe with over 50 contributing standard techniques (Gutiérrez et al., 2018).

Besides these standard SD methods, a number of machine learning techniques have been also adapted and applied for downscaling. For instance, the first applications of neural networks date back to the late 1990s (Wilby et al., 1998; Schoof and Pryor, 2001). More recently, other alternative machine learning approaches have been applied, such as support vector machines (SVMs; Tripathi et al., 2006), random forests (Pour et al., 2016; He et al., 2016) or genetic programming (Sachindra and Kanae, 2019). There have been also a number of intercomparison studies analyzing standard and machine learning techniques (Wilby et al., 1998; Chen et al., 2010; Yang et al., 2016; Sachindra et al., 2018), with an overall consensus that no technique clearly outperforms the others and that limited added value – in terms of performance, interpretability and parsimony – is obtained with sophisticated machine learning options, particularly in the context of climate change studies.

In the last decade, machine learning has gained renewed attention in several fields, boosted by major breakthroughs obtained with deep learning (DL) models (see Schmidhuber, 2015, for an overview). The advantage of DL resides in its ability to extract high-level feature representations in a hierarchical way due to its (deep) layered structure. In particular, in spatiotemporal datasets, convolutional neural networks (CNNs) have gained great attention due to their ability to learn spatial features from data (LeCun and Bengio, 1995). DL models allow high-dimensional problems to be treated automatically, thereby avoiding the use of conventional feature extraction techniques (e.g., principal components, PCs), which are commonly used in more classic approaches (e.g., linear models and traditional fully connected neural networks). Moreover, new efficient learning methods (e.g., batch, stochastic and mini-batch gradient descent), regularization options (e.g., dropout), and computational frameworks (e.g., TensorFlow; see Wang et al., 2019, for an overview) have popularized the use of DL techniques, allowing convolutional neural networks to learn efficiently from (big) data and avoid overfitting. Different configurations of CNNs have proven successful in a variety of problems in several disciplines, particularly in image recognition (Schmidhuber, 2015). There have also been a number of recent successful applications in climate science, including the detection of extreme weather events (Liu et al., 2016), the estimation of cyclone intensity (Pradhan et al., 2018), the detection of atmospheric rivers (Chapman et al., 2019), the emulation of model parameterizations (Gentine et al., 2018; Rasp et al., 2018; Larraondo et al., 2019) and full simplified models (Scher and Messori, 2019). The reader is referred to Reichstein et al. (2019) for a recent overview.

There have been some attempts to test the application of these techniques for SD, including simple illustrative examples of super-resolution approaches to recover high-resolution (precipitation) fields from low-resolution counterparts with promising results (Vandal et al., 2017b; Rodrigues et al., 2018). In the context of SD, deep learning applications have applied complex convolutional-based topologies (Vandal et al., 2017a; Pan et al., 2019), autoencoder architectures (Vandal et al., 2019) and long short-term memory (LSTM) networks (Misra et al., 2018; Miao et al., 2019) over small case study areas and using simple validation frameworks, resulting in different conclusions about their performance, as compared to other standard approaches. Therefore, these complex (in many cases off-the-shelf) models are usually seen as black boxes, generating distrust among the climate community, particularly when it comes to climate change problems. Recently, Reichstein et al. (2019) outlined this problem and encouraged research towards the understanding of deep neural networks in climate science.

In this study we aim to shed light on this problem and perform a comprehensive evaluation of deep SD models of increasing complexity, assessing the particular role of the different elements comprising the deep neural network architecture (e.g., convolutional and fully connected or dense layers). In particular, we use the VALUE validation framework over a continental region (Europe) and compare deep SD methods with a few standard benchmark methods best performing in the VALUE intercomparison (Gutiérrez et al., 2018). Besides this, we also focus on the extrapolation capability of the different methods, which is fundamental for climate change studies. Overall, our results show that simple deep CNNs outperform standard methods (particularly for precipitation) in most of the aspects analyzed.

The code needed to fully replicate the experiments and results shown in this paper is freely available as Jupyter notebooks at the DeepDownscaling GitHub repository (https://github.com/SantanderMetGroup/DeepDownscaling, last access: 23 April 2020; Baño Medina et al., 2020). In addition, in this paper we introduce `downscaleR.keras`, an extension of the `downscaleR` (Bedia et al., 2019) package

that integrates `keras` into the `climate4R` (Iturbide et al., 2019) framework (see the "Code availability" section).

2 Experimental intercomparison framework

2.1 Area of study and data

The VALUE COST Action (2012–2015) developed a framework to validate and intercompare downscaling techniques over Europe, focusing on different aspects such as temporal and spatial structure and extremes (Maraun et al., 2015). The experimental framework for the first experiment (downscaling with "perfect" reanalysis predictors) is publicly available at http://www.value-cost.eu/validation (last access: 23 April 2020) as well as the intercomparison results for over 50 different standard downscaling methods (Gutiérrez et al., 2018). Therefore, VALUE offers a unique opportunity for a rigorous and comprehensive intercomparison of different deep learning topologies for downscaling.

In particular, VALUE proposes the use of 20 standard predictors from the ERA-Interim reanalysis, selected over a European domain (ranging from 36 to 72° in latitude and from −10 to 32° in longitude, with a 2° resolution) for the 30-year period 1979–2008. This predictor set is formed by five large-scale thermodynamic variables (geopotential height, zonal and meridional wind, temperature, and specific humidity) at four different vertical levels (1000, 850, 700 and 500 hPa) each. The left column of Fig. 1 shows the climatology (and the grid) of two illustrative predictors used in this study.

The target predictands considered in this work are surface (daily) mean temperature and accumulated precipitation. Instead of the 86 representative local stations used in VALUE, we used the observational gridded dataset from E-OBS v14 (0.5° resolution). Note that this extended experiment allows for a better comparison with dynamical downscaling experiments carried out under the Coordinated Regional Climate Downscaling Experiment (CORDEX) initiative (Gutowski et al., 2016). The right column of Fig. 1 shows the climatology of the two target predictands: temperature and precipitation.

Daily standardized predictor values are defined considering the closest ERA-Interim grid boxes (one or four) to each E-OBS grid box for the benchmarking linear and generalized linear techniques (see Sect. 2.3). However, the entire domain is used for the deep learning models, which allows testing of their suitability to automatically handle high-dimensional input data, extracting relevant spatial features (note that this is particularly important for continent-wide applications).

2.2 Evaluation indices and cross-validation

The validation of downscaling methods is a multi-faceted problem with different aspects involved, such as the representation of extremes (Hertig et al., 2019) or the temporal (Maraun et al., 2019) and spatial (Widmann et al., 2019) structure. VALUE developed a comprehensive list of indices and measures (available at the VALUE validation portal: http://www.value-cost.eu/validationportal, last access: 23 April 2020) which allows most of these aspects to be properly evaluated. Moreover, an implementation of these indices in an R package (VALUE, https://github.com/SantanderMetGroup/VALUE, last access: 23 April 2020) is available for research reproducibility. In this work we consider the subset of VALUE metrics shown in Table 1 to assess the performance of the downscaling methods to reproduce the observations. Note that different metrics are considered for temperature and precipitation.

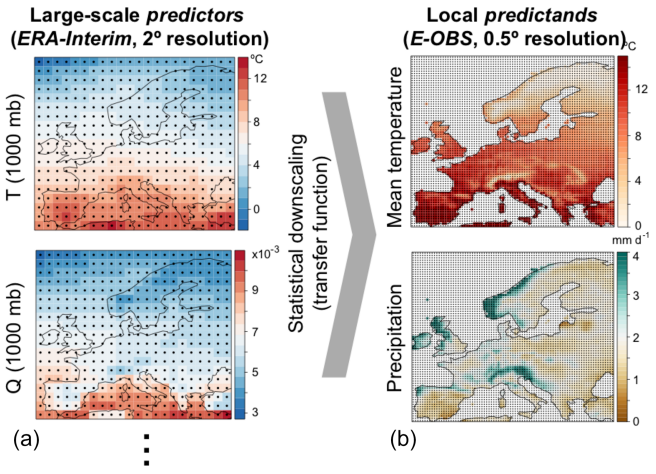

Figure 1. Climatology for **(a)** two typical predictors (air temperature, T, and specific humidity, Q, at 1000 mbar), as given by the ERA-Interim reanalysis (2°), and **(b)** the observed target variables of this work, temperature and precipitation from E-OBS (0.5°). Dots indicate the center of each grid box.

For temperature, biases are given as absolute differences (in °C), whereas for precipitation they are expressed as relative differences with respect to the observed value (in %). Note that, beyond the bias in the mean, we also assess the bias in extreme percentiles, in particular the 2nd percentile (P2, for temperature) and the 98th (P98, for both temperature and precipitation). We also compute the biases for four temporal indices used in VALUE: the median warm (WAMS) and cold (CAMS) annual max spells for temperature and the median wet (WetAMS) and dry (DryAMS) annual max spells for precipitation. In addition to the latter temporal metrics we include the (lag 1) autocorrelation (AC1) for temperatures and the annual cycle's relative amplitude for precipitation, the latter computed as the difference between maximum and minimum values of the annual cycle (defined using a 30 d moving window over calendar days), relative to the mean of these two values. We also consider the root mean square error (RMSE), which measures the average magnitude of the forecast errors; in the case of precipitation this metric is calculated conditioned to observed wet days (rainfall > 1 mm). To evaluate how close the predictions follow the observa-

Table 1. Subset of VALUE metrics used in this study to validate the different downscaling methods considered (see Table 2). The symbol "–" denotes nondimensionality.

Description	Variable	Units
Bias (for the mean)	temp., precip.	°C , %
Bias (for the 2nd percentile, P2)	temp.	°C
Bias (for the 98th percentile, P98)	temp., precip.	°C , %
Root mean square error (RMSE)	temp., precip.	°C, mm d^{-1}
Ratio of standard deviations	temp.	–
Pearson correlation	temp.	–
Spearman correlation	precip.	–
ROC skill score (ROCSS)	precip.	–
Bias (warm annual max spell, WAMS)	temp.	d
Bias (cold annual max spell, CAMS)	temp.	d
Bias (wet annual max spell, WetAMS)	precip.	d
Bias (dry annual max spell, DryAMS)	precip.	d
Bias (lag 1 autocorrelation, AC1)	temp.	–
Bias (relative amplitude of the annual cycle)	precip.	–

tions, we also assess correlation, in particular the Pearson coefficient for temperature and the Spearman rank one (adequate for non-Gaussian variables) for precipitation; for the particular case of temperature, the seasonal cycle is removed from both observations and predictions in order to avoid its (known) effect on the correlation. This is done by removing the annual cycle defined by a 31 d moving window centered on each calendar day. For this variable we also consider the ratio of standard deviations, i.e., that of the predictions divided by that of the observations. Finally, to evaluate how well the probabilistic predictions of rain occurrence discriminate the binary event of rain or no rain, we consider the ROC skill score (ROCSS) (see, e.g., Manzanas et al., 2014), which is based on the area under the ROC curve (see Kharin and Zwiers, 2003, for details).

The VALUE framework builds on a cross-validation approach in which the 30-year period of study (1979–2008) is chronologically split into five consecutive folds. We are particularly interested in analyzing the out-of-sample extrapolation capabilities of the deep SD models. Therefore, following the recommendations of Riley (2019, *"the question you want to answer should affect the way you split your data"*), we focus on the last fold, for which warmer conditions have been observed. Therefore, in this work we apply a simplified hold-out approach using the period 2003–2008 for validation and train the models using the remaining years (1979–2002). Figure 2 shows the climatology of the training period for both temperature and precipitation (top and bottom panel, respectively), as well as the mean differences between the test and the training periods (the latter taken as a reference). For temperature, warmer conditions are observed in the test period – over 0.7° for both mean values and extremes, which is especially significant for the 2nd percentile (cold days), for which temperatures increase up to 2° in northern Europe – compared with the training period. This allows us to estimate the extrapolation capabilities of the different methods, which is particularly relevant for climate change studies.

Importantly, note that the differences between the test and training periods in Fig. 2 reveal some inconsistencies in the dataset for both temperature (southern Iberia and the Alps) and precipitation (northeastern Iberia and the Baltic states). This may be an artifact due to changes or interruptions in the national station networks used to construct E-OBS and may not correspond to a real change in the dataset. This will be taken into account when analyzing the results in Sect. 4.

2.3 Standard statistical downscaling methods used for benchmarking

We use as a benchmark some state-of-the-art standard techniques which ranked among the top in the VALUE intercomparison experiment. In particular, multiple linear and generalized linear regression models (hereafter referred to as GLMs) exhibited good overall performance for temperature and precipitation, respectively (Gutiérrez et al., 2018). Here, we consider the version of these methods described in Bedia et al. (2019) which use the predictor values in the four grid boxes closest to the target location. This choice is a good compromise between feeding the model with full spatial information (all grid boxes, which is problematic due to the resulting high dimensionality) and insufficient spatial representation when considering a single grid box. For the sake of completeness we also illustrate the results obtained with a single grid box, in order to provide an estimate of the added value of extending the spatial information considered for the different variables. These benchmark models are denoted GLM1 and GLM4 for one and four grid boxes, respectively (first two rows in Table 2).

In the case of temperature a single multiple-regression model (i.e., GLM with Gaussian family) is used, whereas

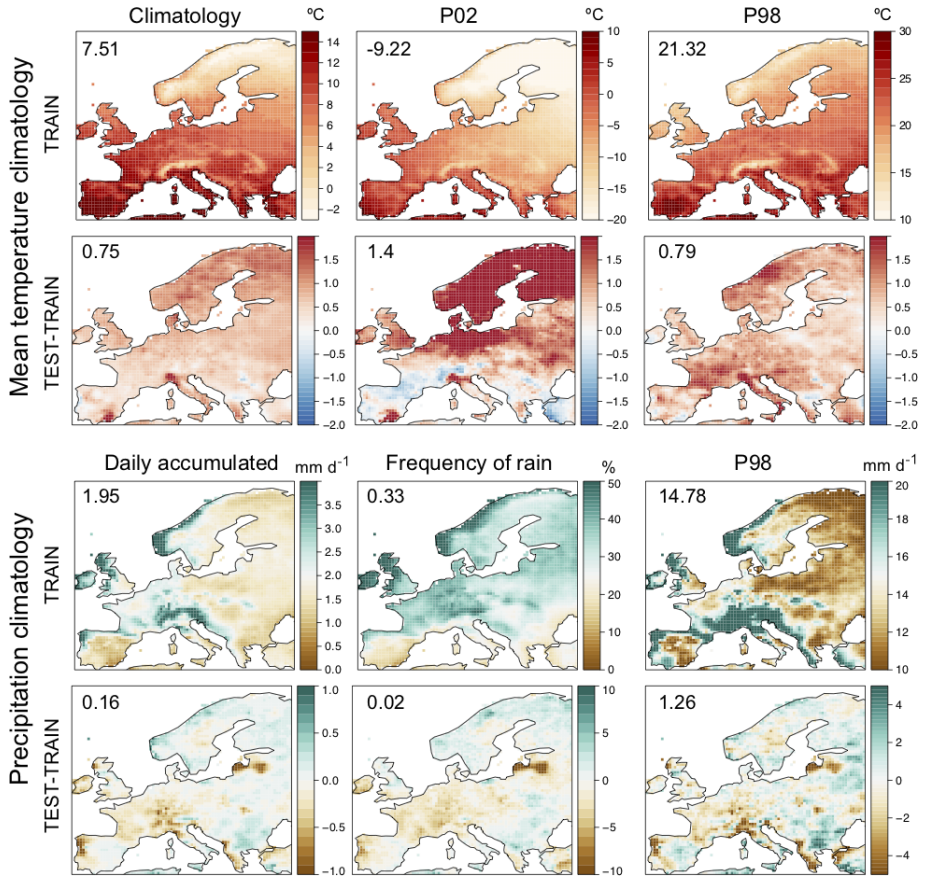

Figure 2. Top panel, top row: E-OBS climatology for the mean value, the P02 and the P98 of temperature in the training period (1979–2002). Top panel, bottom row: mean difference between the test and training periods (the latter taken as a reference) for the different quantities shown in the top row. Bottom panel: as in the top panel but for precipitation, showing the mean value, the frequency of rainy days and the P98. In all cases, the numbers within the panels indicate the spatial mean values.

for precipitation two different GLMs are applied, one for the occurrence (precipitation > 1 mm) and one for the amount of precipitation, using binomial and gamma families with a logarithmic link, respectively (see, e.g., Manzanas et al., 2015). In this case, the values from the two models are multiplied to obtain the final prediction or precipitation, although occurrence and amount are also evaluated separately.

3 Deep convolutional neural networks

Despite the success of deep learning in many fields, these complex and highly nonlinear models are still seen as black boxes, generating distrust among the climate community, particularly when it comes to climate change problems, since their validation and generalization capability is configuration specific and thus difficult to assess in general. Recently, Reichstein et al. (2019) outlined this problem and encouraged research towards the understanding of deep neural networks in climate science. In this study we aim to shed light on the particular role of the different elements comprising the deep neural network architecture (e.g., convolutional and fully connected or dense layers). To do this, we build and evaluate deep SD models of increasing complexity, starting with a simple benchmark linear model (GLM) and adding additional "deep" components, in particular convolution and dense layers, as shown schematically in Fig. 3.

The basic neural network topology relies on feed-forward networks composed of several layers of nonlinear neurons which are fully connected between consecutive layers, from the input to the output (these are commonly referred to as "dense" networks; see Fig. 3). Each of these connections is characterized by a weight which is learned from data (e.g., the two layers of 50 neurons each in Fig. 3 result in a total of 50×50 internal weights, besides the input and output connections). Differently to standard dense networks (whose input is directly the raw predictor data), convolutional networks generate data-driven spatial features to feed the dense network. These layers convolute the raw gridded predictors using 3-D kernels (variable, latitude and longitude), considering a neighborhood of the corresponding grid box (3×3 in this work) in the previous layer (see Fig. 3). Instead of fully

Table 2. Description of the deep learning architectures intercompared in this study, together with the two benchmark methods: GLM1 and GLM4 (these models are trained separately for each of the 3258 land-only grid boxes in E-OBS). Convolutional layers are indicated with boldfaced numbers. The numbers indicating the architecture correspond to the number of neurons in the different layers (in bold for convolutional layers).

Model	Architecture	Rationale
GLM1	20-1 (× 3258)	Simplest linear local model for benchmarking
GLM4	80-1 (× 3258)	Increasing the predictor's spatial domain
CNN-LM	20-**50**-**25**-**1**-3258	Using convolutions to automatically obtain meaningful spatial predictors
CNN1	20-**50**-**25**-**1**-3258	Testing the added value of CNN nonlinearity
CNN10	20-**50**-**25**-**10**-3258	Increasing the complexity of last CNN feature's layer
CNN-PR	20-**10**-**25**-**50**-3258	Using standard topologies from pattern recognition
CNNdense	20-**50**-**25**-**10**-50-50-3258	Using complex dense CNN models

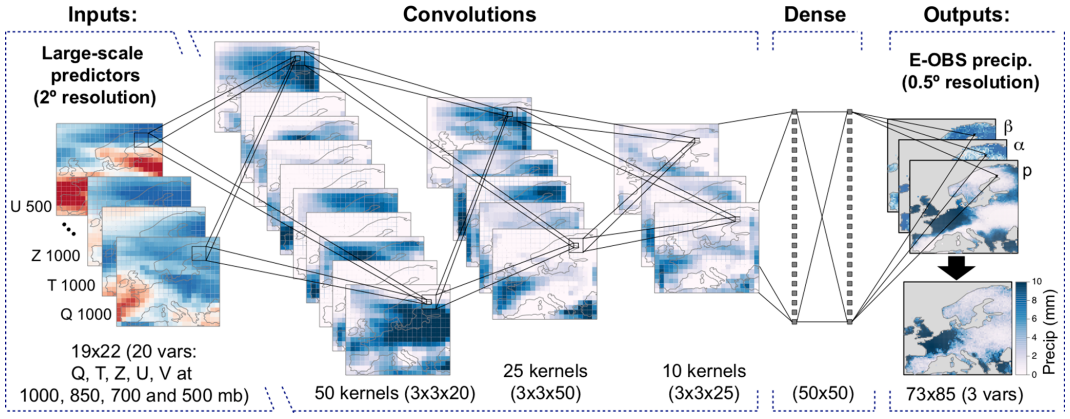

Figure 3. Scheme of the convolutional neural network architecture used in this work to downscale European (E-OBS 0.5° grid) precipitation based on five coarse (2°) large-scale standard predictors (at four pressure levels). The network includes a first block of three convolutional layers with 50, 25 and 10 ($3 \times 3 \times$ no. inputs) kernels, respectively, followed by two fully connected (dense) layers with 50 neurons each. The output is modeled through a mixed binomial–lognormal distribution, and the corresponding parameters are estimated by the network, obtaining precipitation as a final product, either deterministically (the expected value) or stochastically (generating a random value from the predicted distribution). The output layer is activated linearly except for the neurons associated with the parameter p, which present sigmoidal activation functions.

connecting the subsequent layers, kernel weights are shared across regions, resulting in a drastic reduction in the degrees of freedom of the network. Due to these convolutional operations, layers consist of filter maps, which can be interpreted as the spatial representation of the feature learned by the kernel. This is crucial when working with datasets with an underlying spatial structure.

To maximize the performance of convolutional topologies, it is necessary to select an adequate number of layers, number of filter maps and kernel size, which has been done here following a screening procedure testing different configurations varying mainly in the number of layers (up to 6), the kernel size (3×3, 5×5 and 7×7 kernels) and the number of neurons in the dense layer (25, 50 and 100). As a result of this screening we obtained an optimum of three convolutional layers and a 3×3 kernel size; moreover, the best results when including the dense final component were obtained with two layers of 50 neurons each; this resulting configuration is displayed in Fig. 3. Therefore, additional layers seem to not benefit the model due to an over-parameterization when more nonlinearity is actually not needed. Likewise, the final choice of kernel size (3×3) is related to the fact that this is an informative scale for downscaling at the resolution considered in this work, with more spatial information gathered as a result of layer composition. Besides the different deep learning architectures, we also analyzed the effect of basic elements such as the activation function or the layer configuration, testing different configurations.

All the deep models used in this work have been trained using daily data for both predictors and predictand. For temperature, the output is the mean of a Gaussian distribution (one output node for each target grid box) and training is performed by minimizing the mean square error. For precipitation, due to its mixed discrete–continuous nature, the network optimizes the negative log likelihood of a Bernoulli–gamma distribution following the approach previously intro-

duced by Cannon (2008). In particular, the network estimates the parameter p (i.e., probability of rain) of the Bernoulli distribution for rain occurrence and the parameters α (shape) and β (scale) of the gamma rain amount model, as illustrated in the output layer of Fig. 3. The final rainfall value for a given day i, r_i, is then inferred as the expected value of a gamma distribution, given by $r_i = \alpha_i \times \beta_i$.

The first two methods analyzed in this work are the two benchmark GLM models (i.e., multiple linear regression for temperature and Bernoulli–gamma GLM for precipitation) considering local predictors at the nearest (four nearest) neighboring grid boxes. They are labeled as GLM1 (GLM4) in Table 2. Selecting information only from the local grid boxes could be a limitation for the methods, and, therefore, some GLM applications consider spatial features as predictors instead, such as principal components from the empirical orthogonal functions (EOFs) (Gutiérrez et al., 2018). Convolutional networks are automatic feature extraction techniques which learn spatial features of increasing complexity from data in a hierarchical way, due to its (deep) layered structure (LeCun and Bengio, 1995). Therefore, as a third model we test the potential of convolutional layers for spatial feature extraction by considering a linear convolutional neural network with three layers (with 50, 25 and 1 feature each) and linear activation functions (CNN-LM in Table 2). The benefits of nonlinearity are tested considering the same convolutional network, CNN-LM, but with nonlinear (ReLu) activation functions in the hidden layers, making the model nonlinear (CNN1 in Table 2). Moreover, the role of the number of convolutional features in the final layer is tested considering a nonlinear convolutional model, but with 10 feature maps (coded as CNN10).

Note that the previous models are built using a decreasing number of features in the subsequent convolutional layers. However, the approach usually used in computer vision for pattern recognition tasks is the opposite (i.e., the number of convolutional maps increases along the network). Therefore, we also tested this type of architecture considering a convolutional neural network with an increasing number of maps (10, 25 and 50, labeled as CNN-PR).

Finally, a general deep neural network is formed by including a dense (feed-forward) network as an additional block taking input from the convolutional layer (see Fig. 3). This is the typical topology considered in practical applications, which combines both feature extraction and nonlinear modeling capabilities (denoted as CNNdense in Table 2).

All deep learning models listed in Table 2 have been tested with and without padding (padding maintains the original resolution of the predictors throughout the convolutional layers, avoiding the loss of information that may occur near the borders of the domain), keeping in each case the best results for the final intercomparison. Padding was found to be useful only when the amount of feature maps in the last layer was small, so padding is only used for CNN1 model.

4 Results

In this section we intercompare and discuss the performance of the different models shown in Table 2 for temperature (Sect. 4.1) and precipitation (Sect. 4.2).

4.1 Temperature

Figure 4 shows the validation results obtained for temperature in terms of the different metrics explained in Sect. 2.2. Each panel contains seven boxplots, one for each of the methods considered (Table 2), representing the spread of the results along the entire E-OBS grid. In particular, the gray boxes correspond to the 25–75th-percentile range, whereas the whiskers cover the 10–90 % range. The horizontal red line plots the median value obtained from the GLM4 method, which is considered as a benchmark.

In general, all methods provide quite satisfactory results, with low biases and RMSE (panels a, d, e and f), a realistic variability (panel c) and very high correlation values (after removing the annual cycle from the series; panel b). Among the classic linear methods, GLM4 clearly outperforms GLM1, which highlights the fact that including predictor information representative of a wider area around the target point helps to better describe the synoptic features determining the local temperature. However, most of the local variability seems to be explained by linear predictor–predictand relationships, as both GLM4 and CNN-LM provide similar results to more sophisticated neural networks which account for nonlinearity (regardless of their architecture). Nevertheless, the biases provided by CNN1, CNN10, CNN-PR and CNNdense for P02 and P98 are lower than those obtained from GLM1, GLM4 and CNN-LM (panels e and f), which suggests that nonlinearity adds some value to the prediction of extremes. Despite the addition of nonlinearity to the model, benefits of convolutional topologies also include the ability to learn adjustable regions and overcome the restrictive limitation of considering just four neighbors as predictor data. Among the neural-based models, the CNNdense model is the worst in terms of local reproducibility. This suggest that mixing the spatial features learned with the convolutions in dense layers results in a relevant loss of spatial information affecting the downscaling. Furthermore, CNN10 (identified with a darker gray) provides the lowest RMSE and the highest correlations, being overall the best method.

According to the temporal metrics computed (panels g, h and i in Fig. 4) we can state that no method clearly outperforms the others in terms of reproduction of spells for temperature. Despite there being some spatial variability (spread of the boxplots), the median results are nearly unbiased in all cases (except for the CNNdense model).

For a better spatial interpretation of these results, Fig. 5 shows maps for each metric (in columns) for GLM1, GLM4 and CNN10 (in rows), representing the two initial bench-

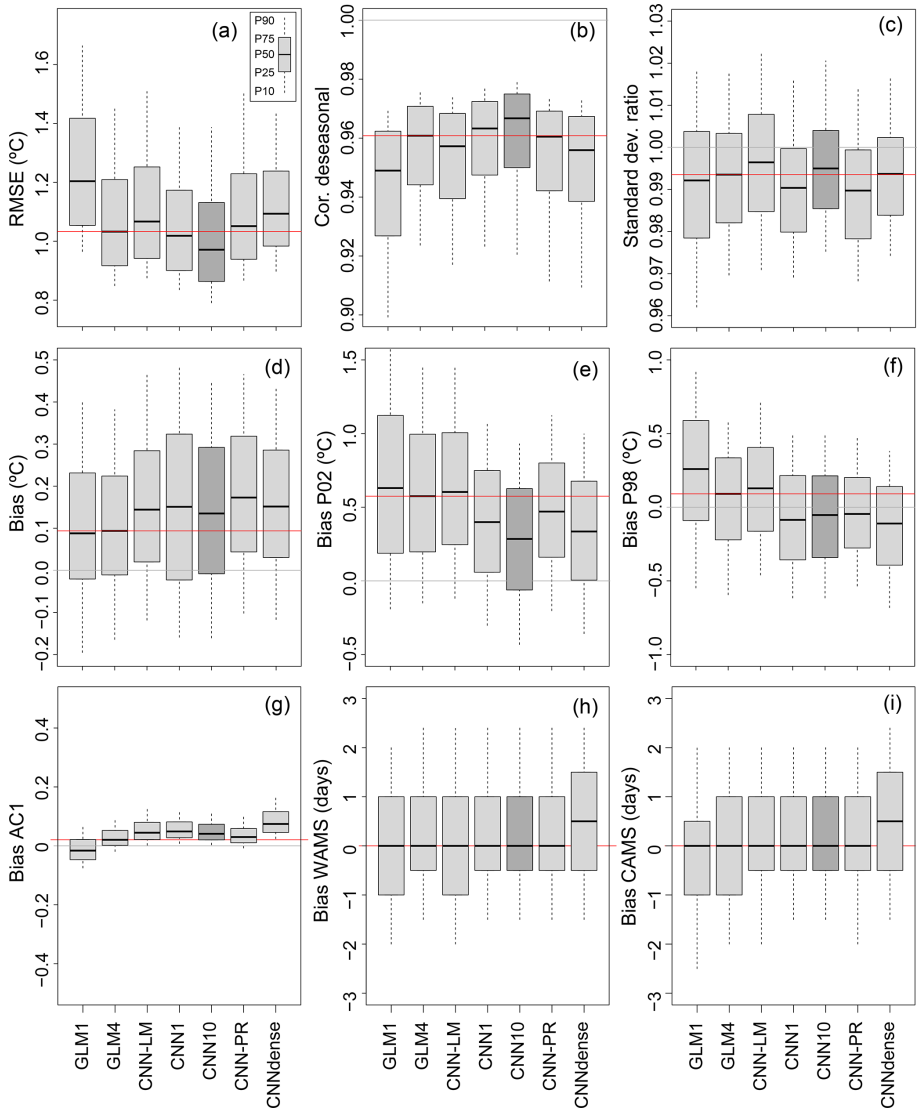

Figure 4. Validation results obtained for temperature. Each panel (corresponding to a particular metric) contains seven boxplots, one for each of the methods tested, which represents the spread of the results along the entire E-OBS grid (the gray boxes correspond to the 25–75th-percentile range, whereas the whiskers cover the 10–90 % range). The horizontal red line plots the median value obtained from the GLM4 method, which is considered as a benchmark, whereas the gray one indicates the "perfect" value for each metric. The dark shaded box indicates the best-performing method, taking into account all metrics simultaneously (CNN10 in this case).

marking methods and the best-performing CNN model in this case. Due to its strong local dependency, GLM1 leads to patchy (discontinuous) spatial patterns, something which is solved by GLM4 – including local predictor information representative of a wider area around the target point provides smother patterns. Beyond this particular aspect, the improvement of GLM4 over GLM1 is evident for RMSE and correlation, and to a lesser extent also for the bias in P98. However, the best results are found for the CNN10 method for the abovementioned particularities, which improves all the validation metrics considered, and in particular the bias for P2. As already pointed out in Sect. 2.1, note that the anomalous results found for southern Iberia could likely be related to issues in the E-OBS dataset.

It is important to highlight that the three methods present very small (mean) biases along the entire continent, which suggests their good extrapolation capability and therefore their potential suitability for climate change studies (recall that the anomalously warm test period that has been selected for this work may serve as a surrogate for the warmer conditions that are expected due to climate change). In order to further explore this issue, we have also analyzed the capability of the models to produce extremes which are larger than those in the calibration data. To this end, we have considered

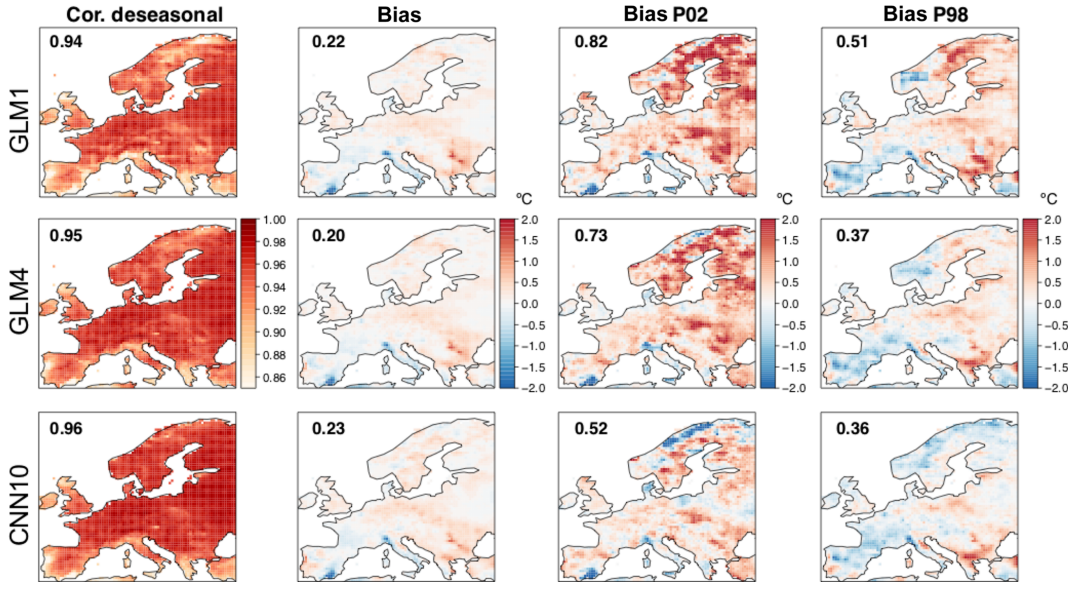

Figure 5. Maps showing the spatial results obtained in terms of the different metrics considered for temperature (in columns) for the two benchmarking versions of GLM (top and middle row) and the best-performing method, the CNN10 (bottom row). The numbers within the panels show the spatial mean absolute values (to avoid error compensation).

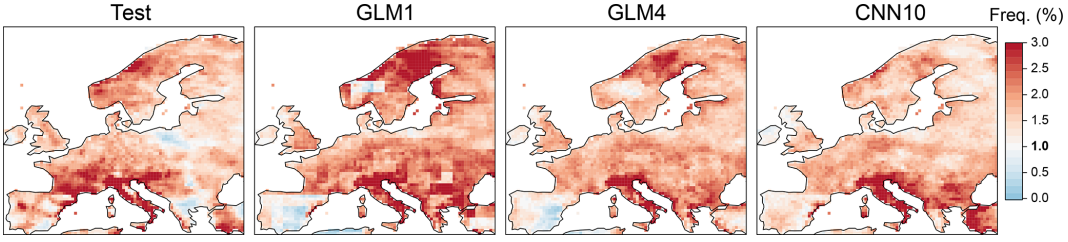

Figure 6. Frequency of exceeding the 99th-percentile value of the training period in each of the grid boxes for the observations in the test period and the test predictions of the GLM1, GLM4 and CNN10 models (in columns). Note that a frequency of 1 % (in boldface) would indicate the same amount of values exceeding the (extreme) threshold as in the training period.

the 99th percentile over the historical period as a robust reference of an extreme value, and calculated the frequency of exceeding this value in the test period for the observations and the GLM1, GLM4 and CNN10 downscaled predictions. The results are shown in Fig. 6 and indicate that the three models (in particular the latter two) are able to reproduce the same frequency and spatial pattern of out-of-sample days observed in the test period.

4.2 Precipitation

Figure 7 is similar to Fig. 4 but for precipitation (note that the validation metrics considered for this variable differ). Similarly to the case of temperature, GLM4 performs notably better than GLM1, in particular for the ROCSS (panel a), the RMSE (panel b) and the correlation (panel c). Note that to compute the ROCSS we use the probabilistic output of the logistic regression for the GLM1 and GLM4 models, and the direct estimation of the parameter p for the neural models. Nevertheless, with the exception of CNN-LM and CNN-PR, convolutional networks yield in general better results than GLM4. Differently to the case of temperature, the results obtained indicate that accounting for nonlinear predictor–predictand relationships is key to better describe precipitation. The latter is based on the improvement of nonlinear models with respect to the linear ones (GLM1, GLM4 and CNN-LM), especially in terms of ROCSS and correlation. Moreover, the standard architecture for pattern recognition (CNN-PR) is not suitable for this prediction problem probably due to an over-parameterization in the connection between the last hidden layer (50 feature maps) and the output layer (three variables per grid point in contrast to the downscaling of temperature where there was only one variable to estimate). In terms of errors (RMSE and the different biases considered), all convolutional networks perform similarly, exhibiting very small biases for the mean centered around zero. With respect to the P98, the slight underestimation shown by deterministic configurations (panel e) can be

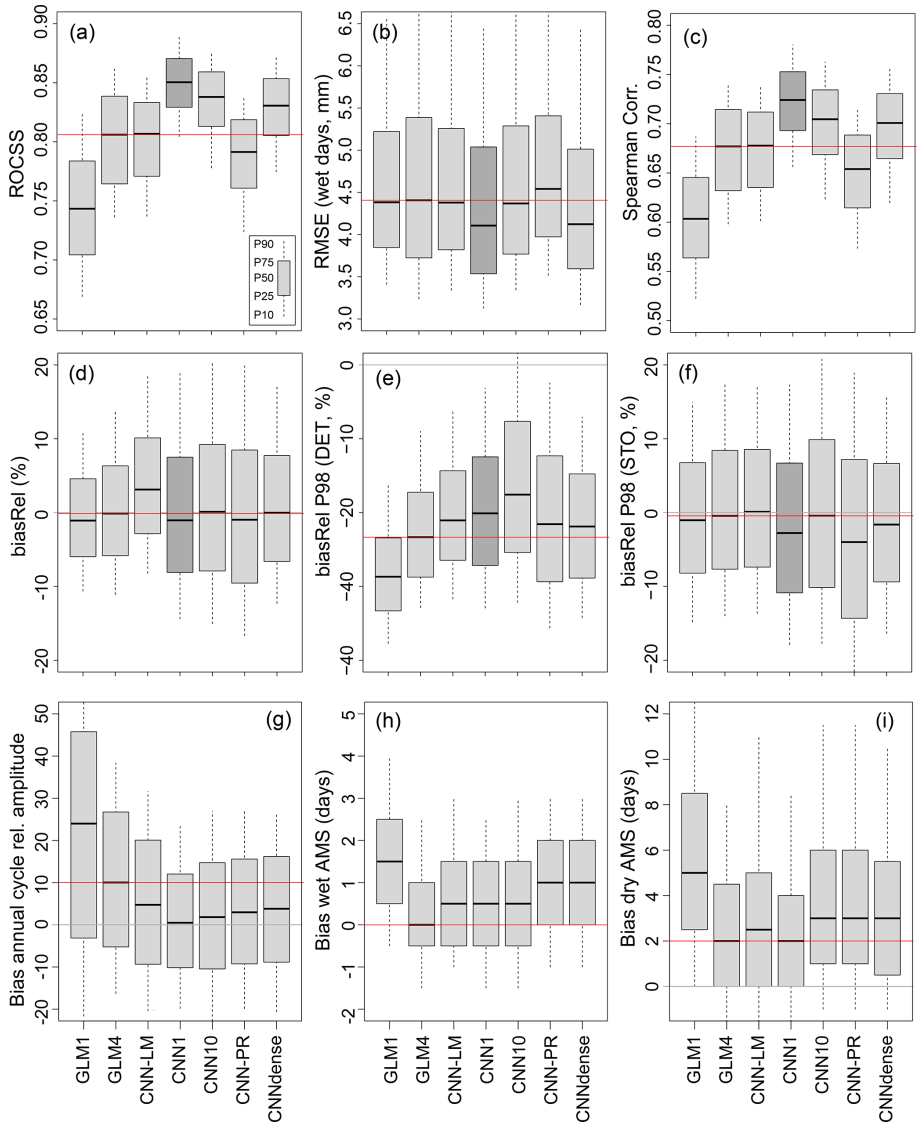

Figure 7. As in Fig. 4 but for precipitation. For the relative bias of the P98 the labels "DET" and "STO" refer to deterministic and stochastic, respectively.

solved by stochastically sampling from the predicted gamma distribution (panel f), but at the cost of losing part of the temporal and spatial correlation achieved by deterministic setups (not shown). Note that, as usual, the correlations found for all methods are much lower than those obtained for temperature, with the CNN-LM method yielding similar values to those obtained with GLM4. The existence of CNN-LM permits marginalizing the role of the convolutions on the spatial predictor data from the nonlinearity of the rest of the neural-based models. This analysis suggests that choosing the four nearest grid boxes as predictors allows the key spatial features that affect the downscaling of precipitation with linear models to be captured (at least over Europe). Differently to the case of temperature, note also that there is not a significant change in the climatological mean between the training and test periods for precipitation (see Fig. 2), so the particular train–test partition considered in this work does not allow a proper assessment of the extrapolation capability of the different methods to be carried out.

Similarly to the analysis of the temperature, there is no clearly outstanding method when analyzing the spells (panels h and i of Fig. 7). GLM4 seems to be unbiased for the WetAMS; however all models tend to overestimate the DryAMS by 2–3 d on average. The GLM1 model performs clearly worse than the rest, probably due to the limited amount of predictor information involved in this method. It has to be noted in this analysis that temporal components have not been explicitly added to the models (e.g., in the form of recurrent connections), whether linear or neural ones, and therefore the reproduction of spells can be affected.

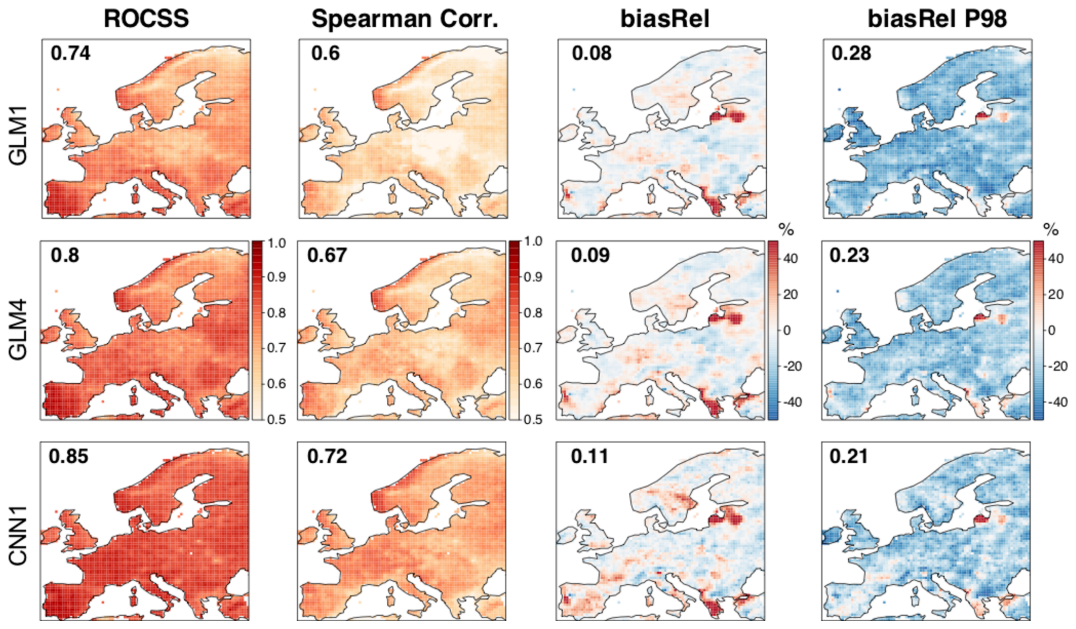

Figure 8. As in Fig. 5 but for precipitation. In this case, CNN1 is taken as the best-performing method (bottom row). The numbers within the panels show the spatial mean absolute values (to avoid error compensation).

Overall, the best results are obtained for CNN1 (marked with a darker gray) and CNNdense, which differ from CNN10 in the amount of neurons placed in the last hidden layer. This suggests that, while one feature map was a little restrictive in the case of temperature, for precipitation 10 maps over-parameterized the network, worsening its generalization capability. The latter may be directly proportional to the number of connections in the output layer, which is dependent on the number of filter maps of the last hidden layer and on the output neurons, which is 3 times bigger for the downscaling of precipitation than for temperature.

Figure 8 is the equivalent of Fig. 5 but for precipitation. Again, the best-performing method (CNN1 in this case; bottom row) is shown, together with the two benchmarking versions of GLM (top and middle rows). In all cases, the deterministic implementation is considered. As for temperature, GLM4 provides better results than GLM1 for all metrics, with the spatial pattern of improvement being rather uniform in all cases. Likewise, CNN1 outperforms GLM4 for all metrics and regions, especially over central and northern Europe. These results suggest the suitability of convolutional neural networks to downscale precipitation, which may be a consequence of their ability to automatically extract the important spatial features determining the local climate, as well as to efficiently model the nonlinearity established between local precipitation and the large-scale atmospheric circulation.

Finally, notice that the anomalous results found over northeastern Iberia and the Baltic states might be due to issues in the E-OBS dataset. Nonetheless, particularly bad results are also found over the Greek peninsula (especially for the mean bias), for which we do not envisage a clear explanation.

5 Conclusions

Deep learning techniques have gained increasing attention due to the promising results obtained in various disciplines. In particular, convolutional neural networks (CNNs) have recently emerged as a promising approach for statistical downscaling in climate due to their ability to learn spatial features from huge spatiotemporal datasets, which would allow for an efficient application of statistical downscaling to large domains (e.g., continents). Within this context, there have been a number of intercomparison studies analyzing standard and machine learning (including CNN) techniques. However, these studies are based on different case studies and use different validation frameworks, which makes a proper assessment of the (possible) added value offered by CNNs difficult and, in some cases, leads to contradictory results (e.g. Vandal et al., 2019; Sachindra et al., 2018).

In this paper we build on a comprehensive framework for validating statistical downscaling techniques (the VALUE validation framework) and evaluate the performance of different CNN models of increasing complexity for downscaling temperature and precipitation over Europe, comparing them with a few standard benchmark methods from VALUE (linear and generalized linear models). Besides analyzing the adequacy of different network architectures, we also focus on their extrapolation capability, a critical point for their possible application in climate change studies, and use a warm test period as a surrogate for possible future climate conditions.

Regarding the classic (generalized) linear methods, our results show that using predictor data in several grid boxes helps to better describe the synoptic features determining the

local climate, thus yielding better predictions both for temperature and precipitation. Furthermore, in the case of temperature, we find that the added value of nonlinear CNNs (regardless of the architecture considered) is limited to the reproduction of extremes, as most of the local variability of this variable is well captured with standard linear methods. However, convolutional topologies can handle high-dimensional domains (i.e., continent-sized) performing an intrinsic feature reduction step in the hidden layers, avoiding tedious and somewhat limited feature selection/reduction techniques out of the learning process. The latter results in an advantage of convolutional networks over classical approaches even when the predictor–predictand link is linear. However, for temperature, mixing the spatial features learned in the dense layers (CNNdense) adds an unnecessary complexity to the network due to the linearity of the link, resulting in worse predictions than those obtained with the GLMs. Moreover, for precipitation, CNNs yield in general better results than standard generalized linear methods, which may reflect the ability of these techniques to automatically extract the important spatial features determining the local climate, as well as to efficiently model the nonlinearity established between this variable and the large-scale atmospheric circulation. In addition, due to the dense connection to the output's layer (which for precipitation is 3 times bigger than for temperature), the size of the last hidden layer plays a major role in the over-parameterization of the net, leading to overfitted predictions when the number of filter maps is too high (e.g., CNN-PR and CNN10). For these reasons, the models CNN1 and CNN10 were found to be the "best" topologies for the downscaling of precipitation and temperature, respectively.

It is worth mentioning that all of the methods considered in this work are specifically designed to reproduce advanced temporal aspects such as spells. In the near future, we plan to explore another battery of methods which explicitly aim to accurately reproduce the observed temporal structure, such as recurrent neural networks.

Note that the overall good results found for the CNNs tested here, together with the fact that they can be suitably applied to large domains without worrying about the spatial features being considered as predictors, can foster their use for statistical downscaling in the framework of international initiatives such as CORDEX, which has traditionally relied on dynamical simulations.

Table A1. Computation times (in minutes) required for the calculation (training and prediction of the test period) for three downscaling methods used in this study: GLM1, GLM4 and CNN1 (the rest of the deep configurations yield similar computing times).

	GLM1	GLM4	CNN1
Precipitation	47	80	74
Temperature	22	28	58

Appendix A: Computing times

In this Appendix we analyze the computation times required for the calculation of the downscaling methods used in this study. All methods build on the R framework `climate4R` (https://github.com/SantanderMetGroup/climate4R, last access: 23 April 2020; Iturbide et al., 2019), in particular on the package `downscaleR` (Bedia et al., 2019) for the linear (GLM) benchmark models and on the package `downscaleR.keras` (presented in this study) for the new deep learning CNN models. In order to test the computational effort of the methods, we have isolated in both packages the code needed to train the models and to predict the test period. The resulting times for both generalized linear models (GLM) and deep CNN models are shown in Table A1, corresponding to the execution on a single machine with the operating system Ubuntu 16.04 LTS (64 bits), with 16 GB of memory and eight Intel® Core™ i7-6700 3.40 GHz processing units.

It must be noted that for precipitation there are two GLMs to train (a binomial logistic and a gamma logarithmic for the occurrence and amount of rain, respectively), and therefore the time included in the table for GLM1 and GLM4 is the sum of these two individual GLMs. Differently, in deep learning models the occurrence and amount of rain are trained simultaneously. In this case, the speed of training depends on some parameters such as the learning rate (learning rate is equal to 0.0001 in this work) and the early-stopping criteria (patience with 30 epochs), which mainly drive the number of epochs or iterations needed to train the model; these parameters have been configured for the particular application of this paper using a screening process.

Table A1 indicates that GLM4 is more time consuming than the simplified counterpart (GLM1) due to a larger number of predictors. Moreover, the time needed to train the deep CNN1 is similar to that required for GLM4 for precipitation (twice for temperature, in agreement with the use of a single model (two models) for temperature (precipitation) GLMs). Therefore, the computational effort is not a strong limitation for continent-wide applications of deep learning models. The main reason for this result is that the GLMs are trained at the grid box level (one model trained for each grid box), whereas the CNN is naturally multisite; therefore, although the training is very time consuming, a single CNN model is needed for the whole domain. However, note that for smaller domains (e.g., nation-wide) the difference between GLMs and CNNs could be large (the computation time of GLMs decreases linearly with the number of grid boxes) and could make a difference.

Code availability. For the purpose of research transparency, we provide notebooks with the full code needed to reproduce the experiments presented in this paper, which can be found in the DeepDownscaling GitHub repository: https://github.com/SantanderMetGroup/DeepDownscaling (last access: 23 April 2020) (Baño Medina et al., 2020). The code builds on the open-source `climate4R` (Iturbide et al., 2019) and `keras` (Chollet, 2015) R frameworks, for the benchmark and the CNN models, respectively. The former is an open R framework for climate data access, processing (e.g., collocation, binding and subsetting), visualization and downscaling (package `downscaleR`; Bedia et al., 2019), allowing for a straightforward application of a wide range of downscaling methods. The latter is a popular R framework for deep learning which builds on `TensorFlow`.

Moreover, in order to facilitate the development of deep learning downscaling methods, we developed an extension of the `downscaleR` package using `keras`, which is referred to as `downscaleR.keras` (https://github.com/SantanderMetGroup/downscaleR.keras, last access: 23 April 2020) and is used for the first time in this paper (see the companion notebooks).

Moreover, the validation of the methods has been carried out with the package `VALUE` and its climate4R wrapper `climate4R.value` (https://github.com/SantanderMetGroup/climate4R.value, last access: 23 April 2020), which enables a direct application of the VALUE validation metrics in the framework of `climate4R`.

Author contributions. JBM and JMG conceived the study. JBM implemented the code to develop the convolutional neural networks and generated the results of the paper. All authors analyzed the results and wrote the manuscript. JBM and RM prepared the code and notebooks for reproducibility.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The authors acknowledge the funding provided by the project MULTI-SDM (CGL2015-66583-R, MINECO/FEDER). They also acknowledge the E-OBS dataset from the EU-FP6 project UERRA (http://www.uerra.eu, last access: 23 April 2020) and the Copernicus Climate Change Service, and the data providers in the ECA&D project (https://www.ecad.eu, last access: 23 April 2020).

Financial support. This research has been supported by the Ministerio de Economía y Competitividad (MULTI-SDM (grant no. CGL2015-66583-R)).

Review statement. This paper was edited by David Topping and reviewed by Matteo De Felice and one anonymous referee.

References

Ba, W., Du, P., Liu, T., Bao, A., Luo, M., Hassan, M., and Qin, C.: Simulating hydrological responses to climate change using dynamic and statistical downscaling methods: a case study in the Kaidu River Basin, Xinjiang, China, J. Arid Land, 10, 905–920, https://doi.org/10.1007/s40333-018-0068-0, 2018.

Baño Medina, J., Manzanas, R., and Gutiérrez, J. M.: SantanderMetGroup/DeepDownscaling: GMD paper accepted for publication (Version v1.2), Zenodo, https://doi.org/10.5281/zenodo.3731351, 2020.

Bedia, J., Baño-Medina, J., Legasa, M. N., Iturbide, M., Manzanas, R., Herrera, S., Casanueva, A., San-Martín, D., Cofiño, A. S., and Gutiérrez, J. M.: Statistical downscaling with the downscaleR package (v3.1.0): contribution to the VALUE intercomparison experiment, Geosci. Model Dev., 13, 1711–1735, https://doi.org/10.5194/gmd-13-1711-2020, 2020.

Cannon, A. J.: Probabilistic Multisite Precipitation Downscaling by an Expanded Bernoulli-Gamma Density Network, J. Hydrometeorol., 9, 1284–1300, https://doi.org/10.1175/2008JHM960.1, 2008.

Chapman, W. E., Subramanian, A. C., Monache, L. D., Xie, S. P., and Ralph, F. M.: Improving Atmospheric River Forecasts With Machine Learning, Geophys. Res. Lett., 46, 10627–10635, https://doi.org/10.1029/2019GL083662, 2019.

Chen, S.-T., Yu, P.-S., and Tang, Y.-H.: Statistical downscaling of daily precipitation using support vector machines and multivariate analysis, J. Hydrol., 385, 13–22, https://doi.org/10.1016/j.jhydrol.2010.01.021, 2010.

Chollet, F.: Keras, available at: https://keras.io (last access: 23 April 2020), 2015.

Galmarini, S., Cannon, A. J., Ceglar, A., Christensen, O. B., de Noblet-Ducoudré, N., Dentener, F., Doblas-Reyes, F. J., Dosio, A., Gutierrez, J. M., Iturbide, M., Jury, M., Lange, S., Loukos, H., Maiorano, A., Maraun, D., McGinnis, S., Nikulin, G., Riccio, A., Sanchez, E., Solazzo, E., Toreti, A., Vrac, M., and Zampieri, M.: Adjusting climate model bias for agricultural impact assessment: How to cut the mustard, Climate Services, 13, 65–69, https://doi.org/10.1016/j.cliser.2019.01.004, 2019.

Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., and Yacalis, G.: Could Machine Learning Break the Convection Parameterization Deadlock?, Geophys. Res. Lett., 45, 5742–5751, https://doi.org/10.1029/2018GL078202, 2018.

Gutiérrez, J. M., San-Martín, D., Brands, S., Manzanas, R., and Herrera, S.: Reassessing Statistical Downscaling Techniques for Their Robust Application under Climate Change Conditions, J. Climate, 26, 171–188, https://doi.org/10.1175/JCLI-D-11-00687.1, 2013.

Gutiérrez, J. M., Maraun, D., Widmann, M., Huth, R., Hertig, E., Benestad, R., Roessler, O., Wibig, J., Wilcke, R., Kotlarski, S., San Martín, D., Herrera, S., Bedia, J., Casanueva, A., Manzanas, R., Iturbide, M., Vrac, M., Dubrovsky, M., Ribalaygua, J., Pórtoles, J., Räty, O., Räisänen, J., Hingray, B., Raynaud, D., Casado, M. J., Ramos, P., Zerenner, T., Turco, M., Bosshard, T., Bartholy, J., Pongracz, R., Keller, D. E., Fischer, A. M., Cardoso, R. M., Soares, P. M. M., Czernecki, B., and Pagé, C.: An intercomparison of a large ensemble of statistical downscaling methods over Europe: Results from the VALUE perfect predictor cross-validation experiment, Int. J. Climatol., 39, 3750–3785, https://doi.org/10.1002/joc.5462, 2018.

Gutowski Jr., W. J., Giorgi, F., Timbal, B., Frigon, A., Jacob, D., Kang, H.-S., Raghavan, K., Lee, B., Lennard, C., Nikulin, G., O'Rourke, E., Rixen, M., Solman, S., Stephenson, T., and Tangang, F.: WCRP COordinated Regional Downscaling EXperiment (CORDEX): a diagnostic MIP for CMIP6, Geosci. Model Dev., 9, 4087–4095, https://doi.org/10.5194/gmd-9-4087-2016, 2016.

He, X., Chaney, N. W., Schleiss, M., and Sheffield, J.: Spatial downscaling of precipitation using adaptable random forests, Water Resour. Res., 52, 8217–8237, https://doi.org/10.1002/2016WR019034, 2016.

Hertig, E., Maraun, D., Bartholy, J., Pongracz, R., Vrac, M., Mares, I., Gutiérrez, J. M., Wibig, J., Casanueva, A., and Soares, P. M. M.: Comparison of statistical downscaling methods with respect to extreme events over Europe: Validation results from the perfect predictor experiment of the COST Action VALUE, Int. J. Climatol., 39, 3846–3867, https://doi.org/10.1002/joc.5469, 2019.

Iturbide, M., Bedia, J., Herrera, S., Baño-Medina, J., Fernández, J., Frías, M. D., Manzanas, R., San-Martín, D., Cimadevilla, E., Cofiño, A. S., and Gutiérrez, J. M.: The R-based climate4R open framework for reproducible climate data access and post-processing, Environ. Model. Softw., 111, 42–54, https://doi.org/10.1016/j.envsoft.2018.09.009, 2019.

Kharin, V. V. and Zwiers, F. W.: On the ROC score of probability forecasts, J. Climate, 16, 4145–4150, https://doi.org/10.1175/1520-0442(2003)016<4145:OTRSOP>2.0.CO;2, 2003.

Larraondo, P. R., Renzullo, L. J., Inza, I., and Lozano, J. A.: A data-driven approach to precipitation parameterizations using convolutional encoder-decoder neural networks, arXiv [physics], arXiv: 1903.10274, 2019.

LeCun, Y. and Bengio, Y.: Convolutional networks for images, speech, and time series, The handbook of brain theory and neural networks, MIT Press, 255–258, 1995.

Liu, Y., Racah, E., Prabhat, Correa, J., Khosrowshahi, A., Lavers, D., Kunkel, K., Wehner, M., and Collins, W.: Application of Deep Convolutional Neural Networks for Detecting Extreme Weather in Climate Datasets, arXiv:1605.01156 [cs], 2016.

Manzanas, R., Frías, M. D., Cofiño, A. S., and Gutiérrez, J. M.: Validation of 40 year multimodel seasonal precipitation forecasts: The role of ENSO on the global skill, J. Geophys. Res.-Atmos., 119, 1708–1719, https://doi.org/10.1002/2013JD020680, 2014.

Manzanas, R., Brands, S., San-Martín, D., Lucero, A., Limbo, C., and Gutiérrez, J. M.: Statistical downscaling in the tropics can be sensitive to reanalysis choice: A case study for precipitation in the Philippines, J. Climate, 28, 4171–4184, https://doi.org/10.1175/JCLI-D-14-00331.1, 2015.

Manzanas, R., Lucero, A., Weisheimer, A., and Gutiérrez, J. M.: Can bias correction and statistical downscaling methods improve the skill of seasonal precipitation forecasts?, Clim. Dynam., 50, 1161–1176, https://doi.org/10.1007/s00382-017-3668-z, 2018.

Maraun, D. and Widmann, M.: Statistical Downscaling and Bias Correction for Climate Research by Douglas Maraun, Cambridge University Press, 2017.

Maraun, D., Widmann, M., Gutiérrez, J. M., Kotlarski, S., Chandler, R. E., Hertig, E., Wibig, J., Huth, R., and Wilcke, R. A.: VALUE: A framework to validate downscaling approaches for climate change studies, Earth's Future, 3, 1–14, https://doi.org/10.1002/2014EF000259, 2015.

Maraun, D., Huth, R., Gutiérrez, J. M., Martín, D. S., Dubrovsky, M., Fischer, A., Hertig, E., Soares, P. M. M., Bartholy, J., Pongracz, R., Widmann, M., Casado, M. J., Ramos, P., and Bedia, J.: The VALUE perfect predictor experiment: Evaluation of temporal variability, Int. J. Climatol., 39, 3786–3818, https://doi.org/10.1002/joc.5222, 2019.

Miao, Q., Pan, B., Wang, H., Hsu, K., and Sorooshian, S.: Improving Monsoon Precipitation Prediction Using Combined Convolutional and Long Short Term Memory Neural Network, Water, 11, 977, https://doi.org/10.3390/w11050977, 2019.

Misra, S., Sarkar, S., and Mitra, P.: Statistical downscaling of precipitation using long short-term memory recurrent neural networks, Theor. Appl. Climatol., 134, 1179–1196, https://doi.org/10.1007/s00704-017-2307-2, 2018.

Pan, B., Hsu, K., AghaKouchak, A., and Sorooshian, S.: Improving Precipitation Estimation Using Convolutional Neural Network, Water Resour. Res., 55, 2301–2321, https://doi.org/10.1029/2018WR024090, 2019.

Pour, S. H., Shahid, S., and Chung, E.-S.: A Hybrid Model for Statistical Downscaling of Daily Rainfall, Procedia Engineer., 154, 1424–1430, https://doi.org/10.1016/j.proeng.2016.07.514, 2016.

Pradhan, R., Aygun, R. S., Maskey, M., Ramachandran, R., and Cecil, D. J.: Tropical Cyclone Intensity Estimation Using a Deep Convolutional Neural Network, IEEE T. Image Process., 27, 692–702, https://doi.org/10.1109/TIP.2017.2766358, 2018.

Rasp, S., Pritchard, M. S., and Gentine, P.: Deep learning to represent subgrid processes in climate models, P. Natl. Acad. Sci. USA, 115, 9684–9689, https://doi.org/10.1073/pnas.1810286115, 2018.

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

Riley, P.: Three pitfalls to avoid in machine learning, Nature, 572, 27–29, https://doi.org/10.1038/d41586-019-02307-y, 2019.

Rodrigues, E. R., Oliveira, I., Cunha, R. L. F., and Netto, M. A. S.: DeepDownscale: a Deep Learning Strategy for High-Resolution Weather Forecast, arXiv:1808.05264 [cs, stat], 2018.

Sachindra, D. A. and Kanae, S.: Machine learning for downscaling: the use of parallel multiple populations in genetic programming, Stoch. Env. Res. Risk A., 33, 1497–1533, https://doi.org/10.1007/s00477-019-01721-y, 2019.

Sachindra, D. A., Ahmed, K., Rashid, M. M., Shahid, S., and Perera, B. J. C.: Statistical downscaling of precipitation using machine learning techniques, Atmos. Res., 212, 240–258, https://doi.org/10.1016/j.atmosres.2018.05.022, 2018.

Sanderson, M., Arbuthnott, K., Kovats, S., Hajat, S., and Falloon, P.: The use of climate information to estimate future mortality from high ambient temperature: A systematic literature review, PLOS one, 12, e0180369, https://doi.org/10.1371/journal.pone.0180369, 2017.

Scher, S. and Messori, G.: Weather and climate forecasting with neural networks: using general circulation models (GCMs) with different complexity as a study ground, Geosci. Model Dev., 12, 2797–2809, https://doi.org/10.5194/gmd-12-2797-2019, 2019.

Schmidhuber, J.: Deep learning in neural networks: An overview, Neural Networks, 61, 85–117, https://doi.org/10.1016/j.neunet.2014.09.003, 2015.

Schoof, J. and Pryor, S.: Downscaling temperature and precipitation: a comparison of regression-based methods and artificial neural networks, Int. J. Climatol., 21, 773–790, https://doi.org/10.1002/joc.655, 2001.

Teutschbein, C., Wetterhall, F., and Seibert, J.: Evaluation of different downscaling techniques for hydrological climate-change impact studies at the catchment scale, Clim. Dynam., 37, 2087–2105, https://doi.org/10.1007/s00382-010-0979-8, 2011.

Tripathi, S., Srinivas, V. V., and Nanjundiah, R. S.: Downscaling of precipitation for climate change scenarios: A support vector machine approach, J. Hydrol., 330, 621–640, https://doi.org/10.1016/j.jhydrol.2006.04.030, 2006.

Vandal, T., Kodra, E., and Ganguly, A. R.: Intercomparison of Machine Learning Methods for Statistical Downscaling: The Case of Daily and Extreme Precipitation, arXiv:1702.04018 [stat], 2017a.

Vandal, T., Kodra, E., Ganguly, S., Michaelis, A., Nemani, R., and Ganguly, A. R.: DeepSD: Generating High Resolution Climate Change Projections through Single Image Super-Resolution, arXiv:1703.03126 [cs.CV], 2017b.

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

Wang, J., Nathan, R., Horne, A., Peel, M. C., Wei, Y., and Langford, J.: Evaluating four downscaling methods for assessment of climate change impact on ecological indicators, Environ. Model. Softw., 96, 68–82, https://doi.org/10.1016/j.envsoft.2017.06.016, 2017.

Wang, Z., Liu, K., Li, J., Zhu, Y., and Zhang, Y.: Various Frameworks and Libraries of Machine Learning and Deep Learning: A Survey, Arch. Computat. Methods Eng., Springer, https://doi.org/10.1007/s11831-018-09312-w, 2019.

Widmann, M., Bedia, J., Gutiérrez, J. M., Bosshard, T., Hertig, E., Maraun, D., Casado, M. J., Ramos, P., Cardoso, R. M., Soares, P. M. M., Ribalaygua, J., Pagé, C., Fischer, A. M., Herrera, S., and Huth, R.: Validation of spatial variability in downscaling results from the VALUE perfect predictor experiment, Int. J. Climatol., 39, 3819–3845, https://doi.org/10.1002/joc.6024, 2019.

Wilby, R. L., Wigley, T. M. L., Conway, D., Jones, P. D., Hewitson, B. C., Main, J., and Wilks, D. S.: Statistical downscaling of general circulation model output: A comparison of methods, Water Resour. Res., 34, 2995–3008, https://doi.org/10.1029/98WR02577, 1998.

Yang, C., Wang, N., Wang, S., and Zhou, L.: Performance comparison of three predictor selection methods for statistical downscaling of daily precipitation, Theor. Appl. Climatol., 131, 43–54, https://doi.org/10.1007/s00704-016-1956-x, 2016.