Assessing Three Perfect Prognosis Methods for Statistical Downscaling of Climate Change Precipitation Scenarios

M N Legasa, S. Thao, M. Vrac, R. Manzanas

Geophysical Research Letters

RESEARCH LETTER
10.1029/2022GL102525

Assessing Three Perfect Prognosis Methods for Statistical Downscaling of Climate Change Precipitation Scenarios

M. N. Legasa[1], S. Thao[2], M. Vrac[2], and R. Manzanas[1,3]

[1]Departamento de Matemática Aplicada y Ciencias de la Computación (MACC), Universidad de Cantabria, Santander, Spain, [2]Laboratoire des Sciences du Climat et de l'Environnement (LSCE-IPSL), CEA/CNRS/UVSQ, Université Paris Saclay, Centre d'Etudes de Saclay, Gif-sur-Yvette, France, [3]Grupo de Meteorología y Computación, Universidad de Cantabria, Unidad Asociada al CSIC, Santander, Spain

Key Points:

- We assess generalized linear models, convolutional neural networks and a posteriori random forests (APRFs) to downscale precipitation from EC-Earth
- The three methodologies lead to plausible local climate change signals, broadly compatible with those given by the raw outputs from EC-Earth
- Moreover, we conclude that APRFs provide the best overall performance, in terms of several metrics

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

M. N. Legasa,
mikel.legasa@unican.es

Citation:

Legasa, M. N., Thao, S., Vrac, M., & Manzanas, R. (2023). Assessing three perfect prognosis methods for statistical downscaling of climate change precipitation scenarios. *Geophysical Research Letters*, *50*, e2022GL102525. https://doi.org/10.1029/2022GL102525

Received 16 DEC 2022
Accepted 23 MAR 2023

Abstract Under the perfect prognosis approach, statistical downscaling methods learn the relationships between large-scale variables from reanalysis and local observational records. These relationships are subsequently applied to downscale future global climate model (GCM) simulations in order to obtain projections for the local region and variables of interest. However, the capability of such methods to produce future climate change signals consistent with those from the GCM, often referred to as *transferability*, is an important issue that remains to be carefully analyzed. Using the EC-Earth GCM and focusing on precipitation, we assess the transferability of generalized linear models, convolutional neural networks and a posteriori random forests (APRFs). We conclude that APRFs present the best overall performance for the historical period, and future local climate change signals consistent with those projected by EC-Earth. Moreover, we show how a slight modification of APRFs can greatly improve the temporal consistency of the downscaled series.

Plain Language Summary Even though they are the main tool to study climate change, global climate models (GCMs) still have a limited spatial resolution (around a hundred kilometers) and exhibit considerable biases with respect to the observed climate. Statistical downscaling aims to solve this issue by learning statistical relationships between large-scale variables, well reproduced by GCMs (e.g., synoptic winds or specific humidity), and local observations of the target surface variable (e.g., precipitation). These relationships are learned over a historical period, and thus a relevant question is whether they can be transferred to the future GCM simulations, that is, whether climate changes produced by GCMs (e.g., changes in mean rainfall) are broadly preserved by the downscaling methods. The rationale behind this is that, even though GCM simulations are biased, GCMs resolve the physical processes responsible for the evolution of the climate system and these changes are thus physically driven. Using the EC-Earth GCM, we assess the transferability of three statistical downscaling methods (generalized linear models, convolutional neural networks and a posteriori random forests (APRFs)) for precipitation downscaling over Europe. We intercompare them using several diagnostic metrics, concluding that APRFs produce reliable projections, with future climate changes consistent with those projected by EC-Earth.

1. Introduction

Global climate models (GCMs) are the main tool used nowadays to simulate the evolution of the climate system at a global scale. Still, when compared to ground-truth observations, they present systematic errors and their spatial resolution—typically around hundreds of kilometers—remains insufficient for most practical applications (see e.g., Doblas-Reyes et al., 2013 and references therein). Besides dynamical downscaling (Giorgi & Mearns, 1999; Laprise et al., 2008; Vaittinada Ayar et al., 2016), statistical downscaling (SD, von Storch et al., 1993) methods aim to alleviate this limitation by statistically linking a set of key large-scale predictors, like geopotential or winds, to the local surface predictand of interest, like precipitation or temperature. Under the perfect prognosis (PP) approach (Bürger & Chen, 2005; Charles et al., 1999; Gutiérrez et al., 2013; Haylock et al., 2006), these empirical/statistical relationships are learned from observed data (some reanalysis is used for the predictors) over a recent historical reference period. Afterward, they are applied to downscale future GCM large-scale predictors, producing projections for the local variable of interest corresponding to the future climate.

Several statistical and machine learning techniques, including analogs (Lorenz, 1969; Zorita & von Storch, 1999), weather generators (Wilby et al., 2002), support vector machines (SVMs) and neural networks (Hastie et al., 2009) have been recently applied to the statistical downscaling of climate change scenarios of different meteorological variables projected by GCMs (Amblar-Francés et al., 2020; Araya-Osses et al., 2020; Baghanam et al., 2020; Fan et al., 2021; Pour et al., 2018; Siabi et al., 2021). Wootten et al. (2020) and Hernanz et al. (2022) performed a comprehensive evaluation of different machine learning alternatives, the latter concluding that the choice of technique can affect the downscaled results up to the point of producing climate change signals of reverse sign for precipitation. Moreover, even for the same SD technique, Manzanas, Fiwa, et al. (2020) showed that the choice of predictor variables considered can also lead to dramatically different precipitation projections.

In this context, Legasa et al. (2022) recently introduced *a posteriori* random forests (APRFs). APRFs are a modification of the random forest (RF) machine learning technique (Breiman, 2001) similar to quantile random forests (Meinshausen, 2006) able to model the whole parametric probability distribution of the target variable. On top of their interpretability and their skill capturing non-linear predictor-predictand relationships, RFs automatically perform feature/predictor selection, thus avoiding the complex, time-consuming and often human-guided task of pre-defining an informative set of predictors. Nevertheless, Legasa et al. (2022) tested APRFs using reanalysis predictors, thoroughly assessing the calibration/training stage, and thus a relevant next question is whether this technique is also suitable for SD of climate change scenarios, that is, using GCM predictors. This issue is often referred to as *transferability* (Dayon et al., 2015; Hernanz et al., 2022). The present article assesses this topic for precipitation, a variable notably difficult to model (see, e.g., Gutiérrez et al., 2019; Legasa et al., 2022) due to its semi-continuous nature (a continuous probability distribution for wet days with positive mass at 0 accounting for dry days) and the limited predictive capability of the large-scale predictors (see Gutiérrez et al., 2019; Themeßl et al., 2011; Vaittinada Ayar et al., 2016 and the references in the next paragraphs).

When producing local future scenarios, it is essential that climate change signals projected by the GCM over the future scenario with respect to the reference historical period (e.g., change in the number of wet days or mean rainfall) are preserved by the statistical downscaling procedure. The rationale behind this is that, even though GCM simulations are biased (Vrac & Friederichs, 2015), GCMs resolve the physical processes that are responsible for the evolution of the climate system and these changes are thus physically driven.

Nevertheless, there is a notable lack of studies focusing on this transferability issue (Baño-Medina et al., 2021; Dayon et al., 2015; Manzanas, Fiwa, et al., 2020). This is particularly the case for RFs, whose most common recent use in literature has been to build a multi-model ensemble of precipitation projections for different representative concentration pathways (RCPs, van Vuuren et al., 2011), as in Ahmed et al. (2020), Homsi et al. (2020), and Sa'adi et al. (2020). RFs were recently compared against other machine learning methodologies in Pham et al. (2019) and Xu et al. (2020). Pham et al. (2019) compared linear discriminant analysis, SVMs and RFs, concluding that RFs outperformed the other methods when used to downscale rainfall discretized in 3 states (dry, non extreme rainfall, extreme rainfall). Xu et al. (2020), instead, assessed three methods (RFs, SVMs, and a deep learning architecture) for downscaling future precipitation under two RCPs, concluding that SVMs were the preferred option. Nevertheless, these two studies used traditional RFs instead of APRFs, which allow us to model the whole distribution of precipitation.

The present article aims to fill this gap of knowledge by assessing the suitability of APRFs to produce local climate change scenarios of precipitation over Europe, using 83 meteorological stations and the RCP8.5 scenario from EC-Earth. Moreover, APRFs are put in context with two other relevant machine learning methodologies: the well-established general linear models (GLMs, Chandler, 2005) and the widely used convolutional neural networks (CNNs, Lecun et al., 1998).

The remainder of this article is structured as follows. In Section 2, we describe the data sets, SD methods and diagnostic metrics we work with. In Section 3, we analyze the results obtained. Lastly, Section 4 summarizes the conclusions drawn from this study.

2. Experimental Framework and Methods

2.1. Data

In this article we follow the experimental framework proposed in the Experiment 2a of the European COST (Cooperation in Science and Technology) action VALUE (Maraun et al., 2015), designed to assess the suitability

of different SD methods to produce local climate change scenarios. We perform downscaling of daily precipitation at the 83 representative locations used in Legasa et al. (2022), which are distributed across Europe (12° West–32° East, 36° North–72° North, see Figure S1 in Supporting Information S1). We build on the PP approach, that is, the three SD techniques considered are trained with ERA-Interim reanalysis (Dee et al., 2011) as large-scale predictors and observed precipitation from ECA&D (European Climate Assessment & Dataset project, Klein Tank et al., 2002) as local-scale predictand. All the methods are trained over the period 1979–2008, and are subsequently applied to downscale precipitation from the EC-Earth GCM, both for the historical and future scenarios.

EC-Earth (Döscher et al., 2022; Hazeleger et al., 2010), a member of the Coupled Model Intercomparison Project Phase 5 (see IPCC, 2014), was chosen for the VALUE Experiment 2a due to its consistency reproducing key large-scale processes affecting the European climate (Lee, 2015). Here, we use both the historical scenario for 1979–2008 (the same period considered to train the three SD methods) and the RCP8.5 (Riahi et al., 2011) for 2071–2100 to analyze the future climate change signals. In all cases, the run *r12i1p1* is considered. The predictors used in this work for both ERA-Interim and EC-Earth are temperature, geopotential, northward wind, eastward wind and specific humidity at 1,000, 850, 700, 500 hPa levels. This selection includes circulation variables, which are less affected by orography and model resolution, together with thermodynamic ones, which are linked to changes in the radiation budget and need to be considered in climate change studies (Huth, 2004).

To avoid the misrepresentation of the annual cycle in the GCM, we corrected the EC-Earth daily predictors, according to ERA-Interim, as follows,

$$\hat{X}_{\mathrm{GCM}} = X_{\mathrm{GCM}} - mean\left(X^{\mathrm{month}}_{\mathrm{HISTORICAL}}\right) + mean\left(X^{\mathrm{month}}_{\mathrm{REANALYSIS}}\right),$$

for the monthly means $mean(X^{\mathrm{month}})$. Note that this simple transformation, applied to both the historical and RCP8.5 predictors, brings the first-order moment of the reanalysis and the GCM into agreement, thereby providing a better approximation for the *perfect prognosis* assumption of relying on predictors well represented by the GCM (Gutiérrez et al., 2019; Manzanas, Fiwa, et al., 2020). EC-Earth was re-gridded from its native spatial resolution (1.12°) to the ERA-Interim's grid considered in VALUE (2°) using bilinear interpolation.

We consider the predictand, precipitation, to follow a Bernoulli-Gamma distribution (Cannon, 2008) parameterized by p (the probability of having a wet day, >1 mm) and α and β (the shape and rate parameters of the Gamma distribution of precipitation amounts on wet days, respectively). The probability density function of the Bernoulli-Gamma distribution, for precipitation y (in mm), is defined as

$$f(y) = \begin{cases} \dfrac{p\beta^{\alpha} y^{\alpha-1} e^{-\beta y}}{\Gamma(\alpha)} & y > 1 \\ 1 - p & 1 >= y >= 0 \end{cases} \quad (1)$$

where $\Gamma(\alpha) = \int_0^{\infty} z^{\alpha-1} e^{-z} dz$ is the Gamma function. The three SD methods used in this work estimate the three parameters of this distribution for each day, whose expected value is $p \cdot \frac{\alpha}{\beta}$.

2.2. Statistical Downscaling Methods

APRFs were introduced and thoroughly assessed in Legasa et al. (2022) for downscaling precipitation intensity under the PP paradigm. APRFs are a modification of traditional random forests that allows for accurately predicting the parametric distribution of any potential variable of interest. In this work we extend the methodology presented in the aforementioned reference, which was originally focused on the Gamma distribution, to model the Bernoulli-Gamma distribution described in the previous section. For this purpose, we update the split function used in Legasa et al. (2022), which is tasked with splitting the predictors' space to provide predictive samples of precipitation, to account for the mixed nature of the Bernoulli-Gamma distribution by considering a mixture of the Gamma deviance and the binary cross-entropy. Specifically, we define the split function to be, for a set of predictive precipitation observations $\{y_i\}$ falling on a leaf,

$$\overbrace{-\overline{p}\log\overline{p}-\left(1-\overline{p}\right)\log\left(1-\overline{p}\right)}^{\text{Bernoulli Entropy}}+2\overbrace{\sum_{y_i^+\in\{y_i\}}\left(-\log\left(\frac{y_i^+}{\overline{y}^+}\right)+\frac{y_i^+-\overline{y}^+}{\overline{y}^+}\right)}^{\text{Gamma Deviance}},$$

where $\overline{p}$ is the proportion of wet days in $\{y_i\}$, y_i^+ is the intensity/rainfall on wet days, and $\overline{y}^+$ the mean precipitation intensity for the wet days. This allows us to estimate, using the a posteriori approach, the three parameters of the Bernoulli-Gamma distribution. Using a cross-validation scheme (not shown for brevity), we selected the optimal configuration of the random forest, corresponding to 200 trees and at least 5 observations in each terminal leaf. The interested reader is referred to Legasa et al. (2022) for further details on the APRF methodology. For each target location, all the gridpoints in the PRUDENCE zone it falls within (see Figure S1 in Supporting Information S1) are used as predictors.

Besides APRFs, two other methodologies which have been used for SD of precipitation in previous studies have been considered in this article. The first one corresponds to the widely used GLMs (see e.g., Chandler and Wheater, 2002), a generalization of traditional linear models which allow for modeling non-normally distributed variables. As done in many previous works (e.g., Abaurrea & Asín, 2005; Manzanas, Fiwa, et al., 2020; Manzanas, Gutiérrez, et al., 2020; Manzanas et al., 2015; Nikulin et al., 2018; San-Martín et al., 2017) we build two independent GLMs: one for modeling precipitation occurrence (p) using the *logit* link and another one for modeling intensity (α/β) using the logarithm link. Note that the latter GLM assumes β to be constant conditional on the predictors' state and is thus estimated from the residuals (see Chandler, 2005). For each target location, both occurrence and intensity GLMs use as predictors the principal components explaining 95% of the variance over the PRUDENCE region it falls within (shown in Figure S1 of the Supporting Information S1). This configuration corresponds exactly to the GLM method used in Gutiérrez et al. (2019) (row 39 in Table 3).

The second one corresponds to a deep learning technique known as convolutional neural network (CNN, Lecun et al., 1998). This methodology was applied to downscale precipitation over E-OBS land-gridpoints in Baño-Medina et al. (2021) in Europe, with the same predictors used in this work. Therefore, we use in the present article the same configuration: the input layer is convolutionally and sequentially connected to 3 hidden layers with 50, 25, and 1 feature maps, with a standard kernel size (3×3) in each convolutional layer. We train the CNNs with Adam optimizer (adaptive moment estimation, Kingma & Ba, 2015), using early stopping with 10% of the data set as validation set. The net is fully connected to the output layer, and, as the APRFs, provides p, α, and β for each day by using the same loss function as in Cannon (2008). The interested reader is referred to Baño-Medina et al. (2020) and Baño-Medina et al. (2021) for further details of this methodology. We use for CNNs the entire geographical domain covering the 83 locations (12° West–32° East, 36° North–72° North) as predictors.

Although we do not show it for the sake of brevity, note that for each of the downscaling methods used in this work we have undertaken a thorough search of the optimal configuration. In addition, we also rely on the conclusions drawn in Baño-Medina et al. (2020) and Legasa et al. (2022), which conducted a comprehensive assessment of the suitability of different settings for CNNs and APRFs for statistical downscaling, respectively. Therefore, the present study provides a representative overview of the merits and demerits of the different techniques considered for our target task.

Finally, note that both CNNs and GLMs require standardization of the predictors, a usual practice in machine learning that avoids issues with the numerical convergence of the algorithms (Hastie et al., 2009). Here, each predictor variable was transformed to have standard deviation 1 and mean 0 by substracting its mean and dividing by its standard deviation at the gridbox level. APRFs do not require this transformation, since the scale of the different predictor variables does not influence the splitting process. In addition, while both APRFs and GLMs build a separate statistical model for each location, CNNs downscale all locations simultaneously with a single model. Using a CNN for each location yielded no significant difference.

2.3. Diagnostic Metrics

To measure the *predictive* performance, in Section 3.1, we use the area under the ROC (*receiver operating characteristic*) curve (AUC, Kharin & Zwiers, 2003); and the Spearman correlation (COR) between the observed and predicted time-series. Note that they are computed for the predicted expected values: for the AUC using p, and for the correlation using $p \cdot \frac{\alpha}{\beta}$.

In addition, a set of diagnostic indicators from the VALUE validation framework has been selected to comprehensively assess the *distributional* performance of the three SD methods considered. R01, SDII, and P98 address the marginal precipitation distribution: R01 measures the proportion of wet (>1 mm/day) days, SDII the mean rainfall on wet days and P98 the 98th percentile of rainfall on wet days, accounting for the tail of the distribution. The remaining indicators focus on temporal aspects. In particular, DW and WW measure the transition probability from wet to dry and from wet to wet days, respectively. DrySpellMean and WetSpellMean, which are only shown in Section 3.4, measure the mean duration of dry and wet spells (≥2 days), respectively. All the indicators are computed from 500 simulations drawn from the downscaled probability distributions. For each particular indicator, these simulations give place to 500 values which are subsequently averaged.

In the next sections, for each indicator we compute (averaged from 500 simulations), when comparing against the reference observed value, we show the relative bias in percentage, computed as $100 \times$ (downscaled − observed)/observed. To assess the climate change signals produced in each diagnostic indicator, we show the relative change in percentage, that is, computed as $100 \times$ (future − historical)/historical.

Note that the standard deviation and the correlation on consecutive wet days were also computed. We found that the standard deviation follows a very similar pattern to P98 in all aspects assessed in this work, and thus we do not show it here for brevity. The correlation for consecutive wet days is very low (maximum observed correlation is 0.32 and median 0.10), and thus we do not assess it in this work.

3. Results

The assessment of the transferability of the three SD methods presented in this work is performed in three steps. First, following a 5-fold cross-validation scheme (Hastie et al., 2009) and using only reanalysis predictors (both for training and predicting, i.e., in *perfect conditions*), we assess the performance of the three methodologies using the AUC, COR, and the marginal distribution indicators described in the previous section (Section 3.1). Second, we apply the SD methodologies, trained using reanalysis for the whole reference historical period, to downscale the historical scenario of the EC-Earth. At this point we aim for the SD methods to provide simulations that reliably reproduce the local observed indicators (Section 3.2). Last, we downscale the RCP8.5 scenario, assessing the consistency between the climate change signals provided by the raw EC-Earth outputs and those downscaled by our three SD methods (Section 3.3). Therefore, in Sections 3.2 and 3.3 the conditions are *non-perfect*, since we apply the relationships learned from reanalysis to the GCM predictors. Section 3.4 is devoted to a small modification of APRFs that leads to better performance in reproducing all the temporal indicators.

3.1. Cross-Validation in *Perfect* Conditions

The same 5-fold cross-validation scheme designed in the Experiment 1a of VALUE and applied in Legasa et al. (2022) was considered to assess the performance of the three SD methods in *perfect* conditions, that is, using ERA-Interim data both for the calibration and the prediction phase. We split the calibration period, 1979–2008, into 5 sets of 6 consecutive years. To predict precipitation for each 6-year period, each statistical model is trained with the remaining 24 years. This way, we assess the performance of the SD methods considered in this work when applied on *unseen* data. The results are shown in Figure 1, which leads to several conclusions.

In terms of predictive performance, measured by the AUC and COR, GLMs fall behind CNNs and APRFs. This difference is not evenly distributed across all locations, with some of them exhibiting significantly poorer performance than others. In particular, the geographical location with worst performance for the GLM (Karasjok, Norway, ECAD ID: 190) has an AUC/COR of 0.74/0.46. APRFs and CNNs significantly outperform GLMs at this location, with an AUC/COR of 0.82/0.51 and 0.81/0.54, respectively. Still, the difference in average predictive power (across all locations) is small: the mean AUC/COR equals 0.886/0.662 for the CNN, 0.886/0.657 for the APRFs and 0.861/0.63 for the GLMs. The differences found between the GLMs and the other two SD methods suggest that capturing non-linear predictor-predictand relationships is particularly relevant at some locations. Moreover, the fact that CNNs and APRFs perform similarly suggests that the convolutional layers over the entire geographical domain do not provide any significant added value.

In terms of distributional performance, the three methodologies reproduce the frequency of wet days reliably, as measured by the R01. Regarding the mean rainfall (SDII), CNNs exhibit some biases, whereas GLMs and APRFs do reproduce it realistically. Indeed, for this measure, the worst bias is −2.67% (−6.58%) for the APRFs (GLMs),

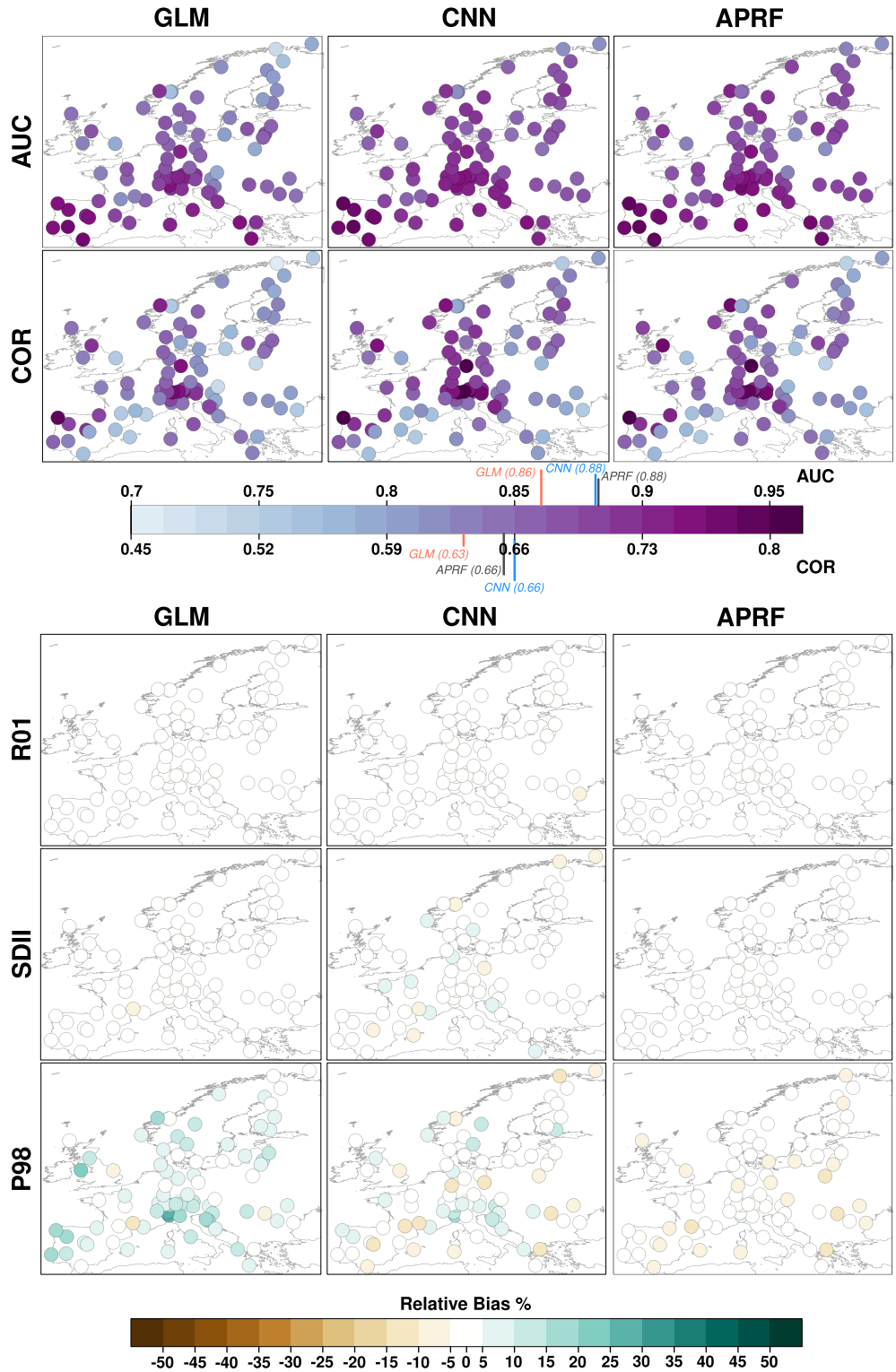

Figure 1. Cross-validated results obtained for the different statistical downscaling methods (in columns) in *perfect* conditions, in terms of some of the diagnostic metrics described in Section 2.3 (in rows). For AUC and COR, the average performance over the 83 stations is shown in the colorbar for each statistical downscaling method.

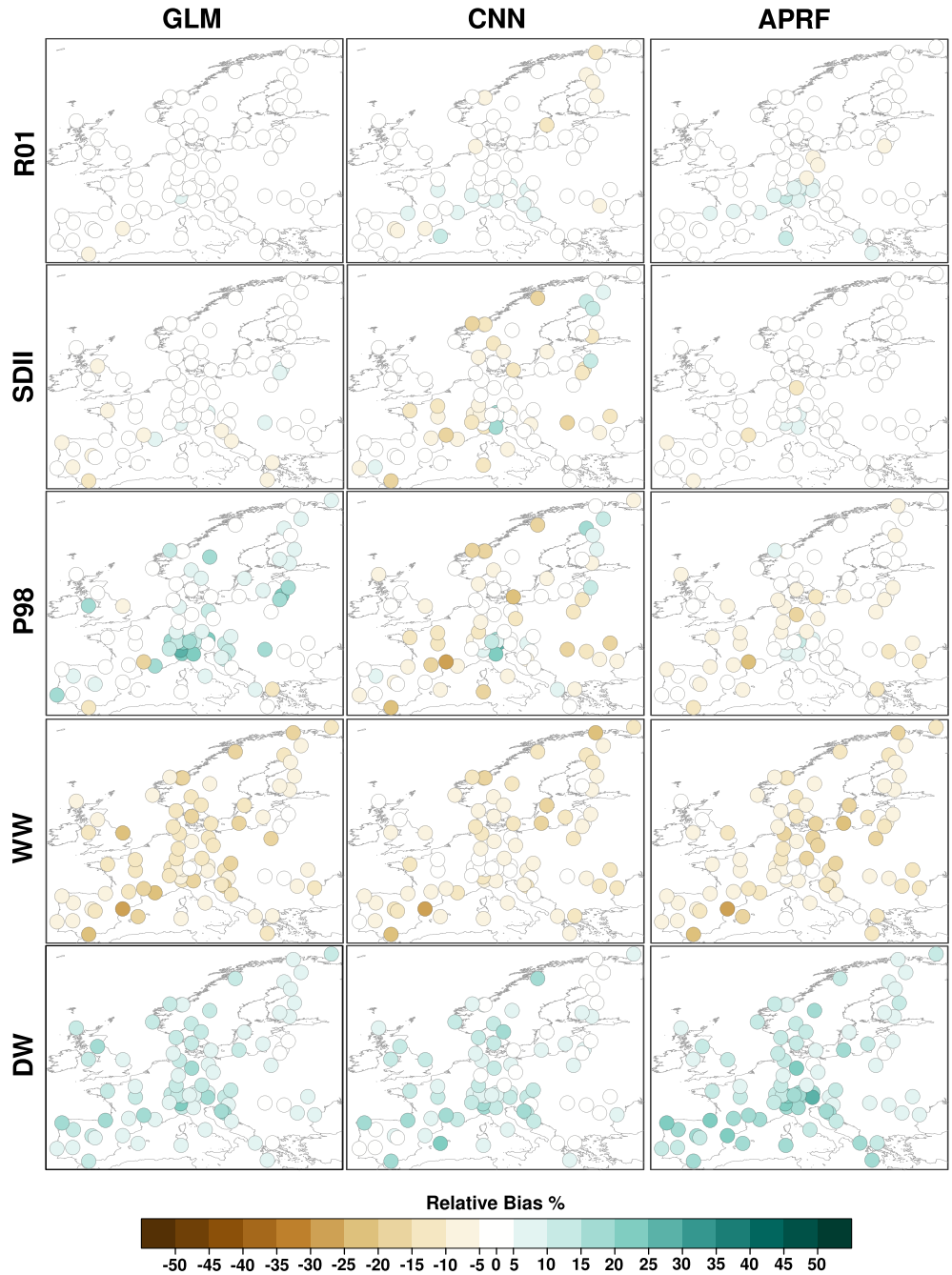

Figure 2. Relative bias with respect to the observations obtained using the different statistical downscaling methods (in columns), computed as 100 × (Downscaled − Observed)/Observed, for the different distributional indicators specified in Section 2.3 (in rows). The methods are trained with reanalysis and applied to downscale the EC-Earth historical scenario.

ranging in between −1.63% and 3.65% for the rest of stations. For the CNNs, 8 locations exhibit a bias in SDII between 5% and 10% and another 8 between −5% and −10%.

All three SD methods suffer from some biases in reproducing the 98th percentile of rainfall, but APRFs perform slightly better than both CNNs and GLMs. On average across the different locations considered, the absolute value of the relative bias (so that negative and positive biases do not compensate) for this indicator is 8.73% for GLMs, 5.94% for CNNs and 3.78% for APRFs. Therefore, taking into account R01, SDII, and P98, we conclude that APRFs provide the best results in terms of distributional performance.

3.2. Downscaling in *Non-Perfect* Conditions: EC-Earth Historical Scenario

We assess in this section the three SD methods in *non-perfect* conditions. This means that they are trained with reanalysis predictors (for the period 1979–2008) and are subsequently applied to downscale the historical scenario of the EC-Earth for the same period (1979–2008). Ideally, we want them to provide simulations that reliably reproduce the observed statistics at the different locations, correcting the bias exhibited by the EC-Earth at the nearest gridpoint (shown in Figure S2 of the Supporting Information S1). The results obtained are shown in Figure 2.

As opposed to the considerably good performance shown in perfect conditions (Section 3.1), the three SD methods exhibit more bias when applied to the EC-Earth predictors. Although relatively weak for R01, the effect is particularly noticeable for SDII, which reaches more than 25% relative bias for the CNN at some locations, with APRFs and GLMs leading to better, similar results. Moreover, APRFs also perform better than GLMs and CNNs at capturing the P98, which is generally overestimated by the GLMs and underestimated by the CNNs. The transitions (WW and DW) are not well reproduced by any of the three SD methods. On average across the different locations considered, GLMs, CNNs, and APRFs underestimate WW by −11%, −8.2%, and −10.2% and overestimate DW by 9.7%, 8.3%, and 12.9%, respectively. This is the reason motivating the introduction of TAPRFs, a small modification to APRFs which improves this aspect, described in Section 3.4.

3.3. Downscaling EC-Earth RCP8.5 Scenario

In this section we address the climate change signals produced by each SD method when downscaling the historical and future (RCP8.5) EC-Earth scenarios. To do so, we compare the local downscaled relative signals against those produced by the raw EC-Earth outputs at the nearest gridpoint, thus assessing whether the changes produced by the different methodologies are consistent both with EC-Earth and among them. Figure 3 shows the EC-Earth projected changes for 2071–2100 under the RCP8.5 emissions scenario for Europe, along with the comparison of the projected/downscaled changes for the different indicators and SD methods used in this work.

Overall, the three SD methods considered yield local climate change signals which are broadly compatible with those given by the raw outputs from EC-Earth. However, GLMs and CNNs tend to deviate more from the changes projected by EC-Earth than APRFs. This effect is particularly evident for GLMs and the P98 indicator, with a clear amplification of the changes expected. Quantitatively, the average difference (across all locations and measured as the absolute value of the relative changes to avoid negative and positive values compensating) between the EC-Earth's raw climate change signal and the signal downscaled by GLMs/CNNs/APRFs is 6.52/4.79/3.54 for the R01; 7/4.93/4.49 for the SDII and 10.1/7.8/7.73 for the P98. For the transitions, these differences are 11.66/10.47/6.38 in WW and 4.39/4.7/5.98 in DW.

A potential explanation for these results might be related to how the different SD methodologies extrapolate values outside the historical domain: GLMs extrapolate exponentially due to the logarithm link, and CNNs response is highly non-linear, by construction. APRFs, instead, estimate distributions from predictive historical records, thus leading to more constrained climate change signals.

3.4. Temporal A Posteriori Random Forests

Figure 2 shows that none of the downscaling methodologies is able to accurately reproduce the observed transitions, neither WW nor DW. Although we do not show it for brevity, this occurs not only when downscaling the historical scenario of EC-Earth, but also in perfect conditions. A straightforward way to overcome this limitation is by introducing a small variation to APRFs: instead of estimating the probability p of a wet day, the already-trained APRF can be used to produce downscaled estimates of the probability of precipitation for each day conditional on the state of the previous day. That is, instead of estimating the probability of precipitation over 1 mm, p, we now estimate p_{DW} and p_{WW}, with $p_{DW} = Probability(\text{precipitation}(t) > 1 \text{ mm} \mid \text{precipitation}(t-1) \leq 1 \text{ mm})$ and $p_{WW} = Probability(\text{precipitation}(t) > 1 \text{ mm} \mid \text{precipitation}(t-1) > 1 \text{ mm})$, where precipitation($t$) and precipitation($t-1$) indicate precipitation on day t and $t-1$, respectively. We can then simulate precipitation on day t by taking into account if day $t-1$ was wet or dry. We call this method *temporal* a posteriori random forest (TAPRF).

Results for the TAPRFs, compared against the APRFs, are shown in Figure 4. Note that only the temporal metrics in Section 2.3 are shown, since both APRFs and TAPRFs perform similarly for the rest of diagnostic metrics.

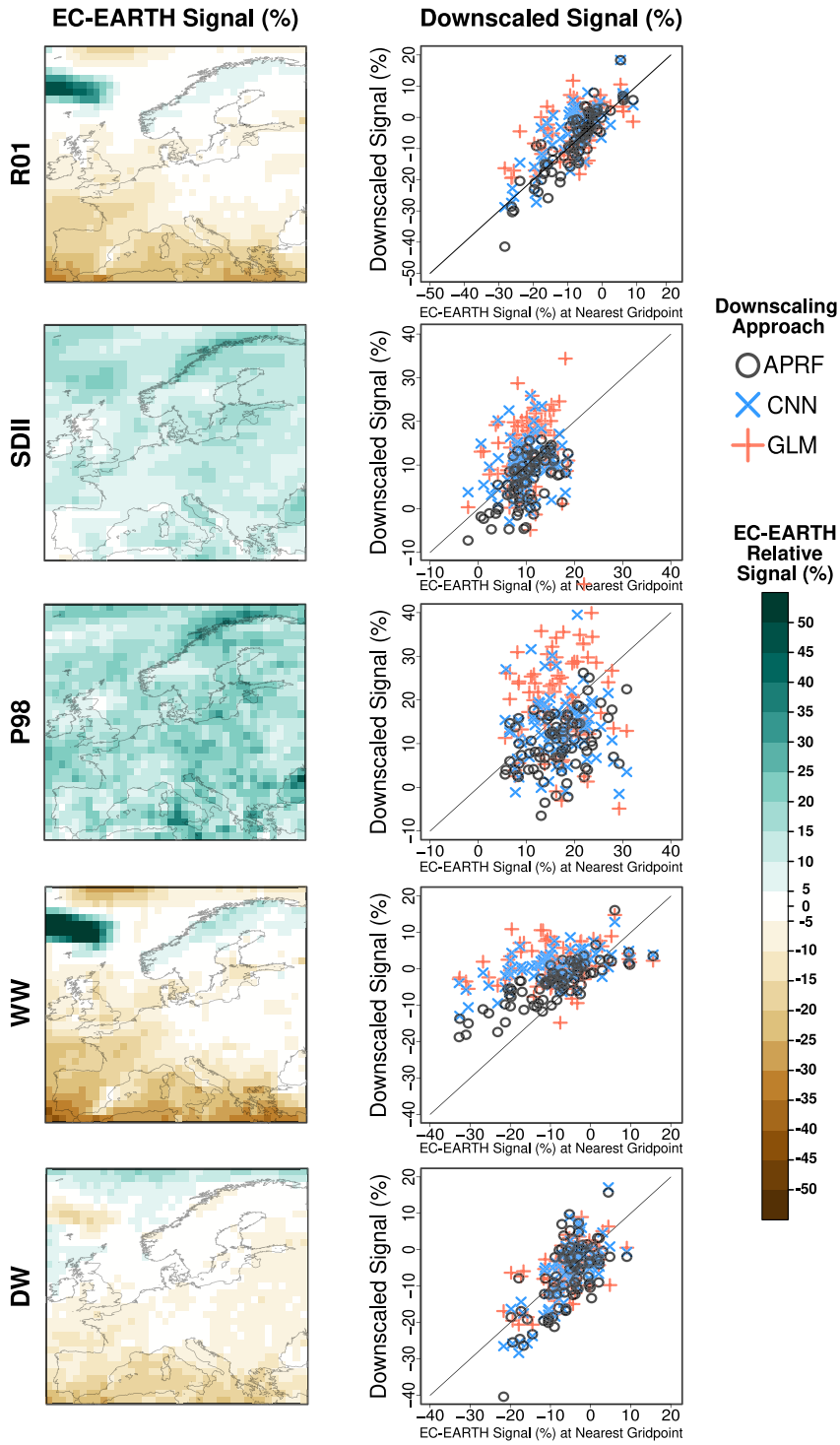

Figure 3. The left column shows the climate change signals projected by the EC-Earth for the period 2071–2100 under the RCP8.5 scenario, expressed as relative changes (%) across the entire Europe. The right column shows the climate change signals produced by each downscaling methodology (*y*-axis) compared against the changes produced by the EC-Earth at the nearest gridpoint (*x*-axis). These signals are computed in percentage as 100 × (Future − Historical)/Historical, for the indicators specified in Section 2.3.

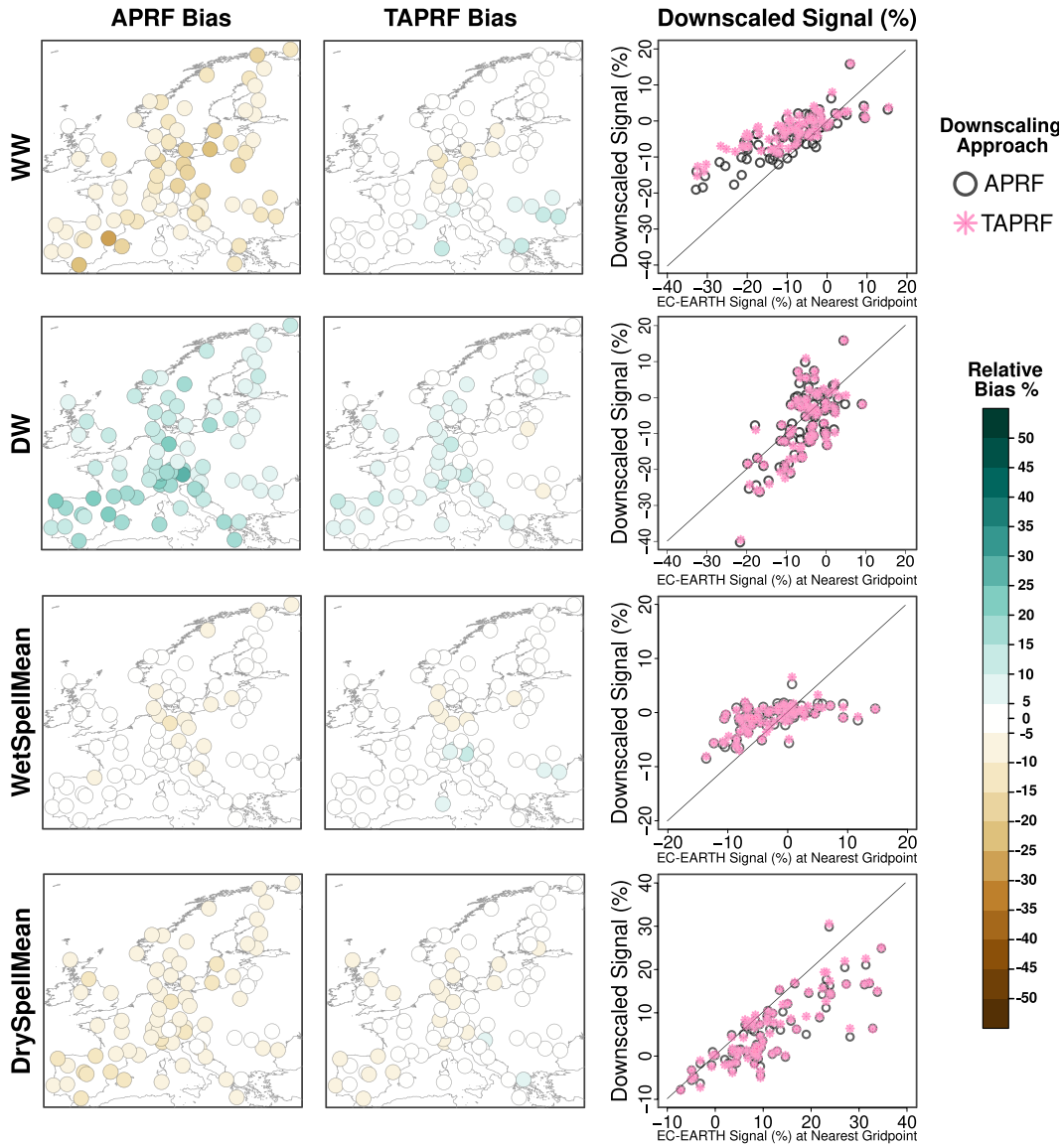

Figure 4. For the temporal metrics described in Section 2.3, relative bias with respect to the observations for the historical scenario, for the APRFs (first column) and TAPRFs (second column). The rightmost column shows, with different symbols (see the legend) the climate change signals produced by APRFs and TAPRFs (*y*-axis) compared against the signals produced by EC-Earth at the closest gridpoint (*x*-axis) for the 2071–2100 period under the RCP85 scenario.

Although the two configurations lead to similar climate change signals (right column), it can be clearly seen that the TAPRFs provide much better results than APRFs in terms of bias for all the temporal indicators (left and middle columns): the average (across locations) absolute value of the relative bias is reduced from 10.3%, 12.9%, 3.1%, and 7.6% for the APRF to 3.8%, 4.7%, 2.7%, and 3.7% for the TAPRF, for the indicators WW, DW, WetSpellMean and DrySpellMean, respectively.

This suggests that TAPRFs are preferable over APRFs for the generation of local climate change scenarios, since they are expected to provide more reliable projections in terms of temporal structure, which is relevant in impact sectors such as hydrology and agriculture.

4. Conclusions

This work presents a comprehensive assessment of the suitability of three perfect prognosis methods (GLMs, CNNs, and APRFs) for statistical downscaling of climate change precipitation scenarios at 83 locations

distributed across Europe. Under the PP assumption, we focus on the *transferability* issue, that is, on whether the relationships learned using reanalysis predictors and observed records can be transferred to downscale climate change scenarios projected by GCMs.

Both APRFs and CNNs provide better predictive performance, as measured by the higher correlation with the observed series and AUC in perfect conditions, suggesting that capturing non-linear predictor-predictand relationships is relevant for some geographical locations. In addition, these two SD methods automatically extract the relevant information contained in the predictor fields, avoiding the need to conduct an exhaustive predictor/geographical domain screening, a complex, time-consuming task that has to be typically undertaken in many widely used statistical downscaling techniques and, in particular, in GLMs (Manzanas, Fiwa, et al., 2020).

In general, the three machine learning methodologies tested lead to local climate change signals which are broadly compatible with those given by the raw outputs from EC-Earth. Nevertheless, GLMs and CNNs tend to deviate more from the changes projected by EC-Earth at some locations than APRFs, sometimes considerably amplifying the changes. APRFs, instead, yield more stable results, which are in better agreement with those projected by EC-Earth, while also performing slightly better than CNNs and GLMs in terms of bias when downscaling the historical scenario of the EC-Earth. Moreover, a slight modification of ARPFs that explicitly models the transition probabilities from dry/wet to wet days, which we call TAPRF, greatly improves the performance in all the temporal indicators, giving higher confidence about the plausibility of the local scenarios obtained with this technique. This modification also illustrates the extensibility of the technique, which we plan to apply to other and more complex distributions, including to perform multivariable downscaling, for example, to simultaneously produce consistent scenarios of precipitation and temperature; and to other GCMs and regions.

Taking into account that the main limitation of any machine learning method is its limited capability to extrapolate outside the predictor values' range, the results presented in this article are promising. Nevertheless, even though the climate change signals are overall consistent with those projected by the EC-Earth, a machine learning methodology with a properly controlled extrapolation mechanism of the climate change signals is still lacking and is an important research perspective.

Data Availability Statement

All the data used in this work (ECA&D observed rainfall, ERA-Interim and EC-Earth predictors), are publicly available and can be downloaded from http://www.value-cost.eu/data. The R (R Core Team, 2020) packages *downscaleR* (Bedia et al. (2020), https://github.com/SantanderMetGroup/downscaleR), *downscaleR.keras* (Baño-Medina et al. (2021), https://github.com/SantanderMetGroup/downscaleR.keras) and *RandomForestDist* (Legasa et al. (2022), https://github.com/MNLR/RandomForestDist) were used, respectively, to train the generalized linear models, convolutional neural networks and a posteriori random forests used in this work.

Acknowledgments
This study is part of the R&D project "*Eventos extremos compuestos para la evaluación de los impactos del cambio climático en la agricultura"* (*COMPOUND*: *TED2021-131334A-I00*) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. R. Manzanas acknowledges support from the R&D project *"Contribución a la nueva generación de proyecciones climáticas regionales de CORDEX mediante técnicas dinámicas y estadísticas"* (CORDyS: PID2020-116595RB-I00). M. Vrac and S. Thao acknowledge support from the H2020 funded project XAIDA with the Grant Agreement number 101003469, and from the COESION project funded by the French National program LEFE (*Les Enveloppes Fluides et l'Environnement*). Additionally, M. N. Legasa acknowledges partial funding by the French embassy in Spain (*"Convocatoria de proyectos científicos de la Embajada de Francia en España para el año 2022"*).

References

Abaurrea, J., & Asín, J. (2005). Forecasting local daily precipitation patterns in a climate change scenario. *Climate Research*, *28*(3), 183–197. https://doi.org/10.3354/cr028183

Ahmed, K., Iqbal, Z., Khan, N., Rasheed, B., Nawaz, N., Malik, I., & Noor, M. (2020). Quantitative assessment of precipitation changes under CMIP5 RCP scenarios over the northern sub-Himalayan region of Pakistan. *Environment, Development and Sustainability*, *22*(8), 7831–7845. https://doi.org/10.1007/s10668-019-00548-5

Amblar-Francés, M. P., Ramos-Calzado, P., Sanchis-Lladó, J., Hernanz-Lázaro, A., Peral-García, M. C., Navascués, B., et al. (2020). High resolution climate change projections for the Pyrenees region. *Advances in Science and Research*, *17*, 191–208. https://doi.org/10.5194/asr-17-191-2020

Araya-Osses, D., Casanueva, A., Román-Figueroa, C., Uribe, J. M., & Paneque, M. (2020). Climate change projections of temperature and precipitation in Chile based on statistical downscaling. *Climate Dynamics*, *54*(9–10), 4309–4330. https://doi.org/10.1007/s00382-020-05231-4

Baghanam, A. H., Nourani, V., Sheikhbabaei, A., & Seifi, A. J. (2020). Statistical downscaling and projection of future temperature change for Tabriz city, Iran. *IOP Conference Series: Earth and Environmental Science*, *491*(1), 012009. https://doi.org/10.1088/1755-1315/491/1/012009

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

Baño-Medina, J., Manzanas, R., & Gutiérrez, J. M. (2021). On the suitability of deep convolutional neural networks for continental-wide downscaling of climate change projections. *Climate Dynamics*, *57*(11), 2941–2951. https://doi.org/10.1007/s00382-021-05847-0

Bedia, J., Baño Medina, J., Legasa, M. N., Iturbide, M., Manzanas, R., Herrera, S., et al. (2020). Statistical downscaling with the downscaleR package (v3.1.0): Contribution to the VALUE intercomparison experiment. *Geoscientific Model Development*, *13*(3), 1711–1735. https://doi.org/10.5194/gmd-13-1711-2020

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bürger, G., & Chen, Y. (2005). Regression-based downscaling of spatial variability for hydrologic applications. *Journal of Hydrology*, *311*(1–4), 299–317. https://doi.org/10.1016/j.jhydrol.2005.01.025

Cannon, A. J. (2008). Probabilistic multisite precipitation downscaling by an expanded Bernoulli–Gamma density network. *Journal of Hydrometeorology*, *9*(6), 1284–1300. https://doi.org/10.1175/2008JHM960.1

Chandler, R. E. (2005). On the use of generalized linear models for interpreting climate variability. *Environmetrics*, *16*(7), 699–715. https://doi.org/10.1002/env.731

Chandler, R. E., & Wheater, H. S. (2002). Analysis of rainfall variability using generalized linear models: A case study from the west of Ireland. *Water Resources Research*, *38*(10), 10–1–10–11. https://doi.org/10.1029/2001WR000906

Charles, S. P., Bates, B. C., & Hughes, J. P. (1999). A spatiotemporal model for downscaling precipitation occurrence and amounts. *Journal of Geophysical Research: Atmospheres*, *104*(D24), 31657–31669. https://doi.org/10.1029/1999JD900119

Dayon, G., Boé, J., & Martin, E. (2015). Transferability in the future climate of a statistical downscaling method for precipitation in France. *Journal of Geophysical Research: Atmospheres*, *120*(3), 1023–1043. https://doi.org/10.1002/2014JD022236

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., et al. (2011). The ERA-interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, *137*(656), 553–597. https://doi.org/10.1002/qj.828

Doblas-Reyes, F. J., García-Serrano, J., Lienert, F., Biescas, A. P., & Rodrigues, L. R. L. (2013). Seasonal climate predictability and forecasting: Status and prospects. *WIREs Climate Change*, *4*(4), 245–268. https://doi.org/10.1002/wcc.217

Döscher, R., Acosta, M., Alessandri, A., Anthoni, P., Arsouze, T., Bergman, T., et al. (2022). The EC-Earth3 Earth system model for the coupled model intercomparison project 6. *Geoscientific Model Development*, *15*(7), 2973–3020. https://doi.org/10.5194/gmd-15-2973-2022

Fan, X., Jiang, L., & Gou, J. (2021). Statistical downscaling and projection of future temperatures across the Loess Plateau, China. *Weather and Climate Extremes*, *32*, 100328. https://doi.org/10.1016/j.wace.2021.100328

Giorgi, F., & Mearns, L. O. (1999). Introduction to special section: Regional climate modeling revisited. *Journal of Geophysical Research*, *104*(D6), 6335–6352. https://doi.org/10.1029/98JD02072

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

Gutiérrez, J. M., San-Martín, D., Brands, S., Manzanas, R., & Herrera, S. (2013). Reassessing statistical downscaling techniques for their robust application under climate change conditions. *Journal of Climate*, *26*(1), 171–188. https://doi.org/10.1175/JCLI-D-11-00687.1

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer New York. https://doi.org/10.1007/978-0-387-84858-7

Haylock, M. R., Cawley, G. C., Harpham, C., Wilby, R. L., & Goodess, C. M. (2006). Downscaling heavy precipitation over the United Kingdom: A comparison of dynamical and statistical methods and their future scenarios. *International Journal of Climatology*, *26*(10), 1397–1415. https://doi.org/10.1002/joc.1318

Hazeleger, W., Severijns, C., Semmler, T., Ştefănescu, S., Yang, S., Wang, X., et al. (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

Hernanz, A., García-Valero, J. A., Domínguez, M., & Rodríguez-Camino, E. (2022). Evaluation of statistical downscaling methods for climate change projections over Spain: Future conditions with pseudo reality (transferability experiment). *International Journal of Climatology*, *42*(7), 3987–4000. https://doi.org/10.1002/joc.7464

Homsi, R., Shiru, M. S., Shahid, S., Ismail, T., Harun, S. B., Al-Ansari, N., et al. (2020). Precipitation projection using a CMIP5 GCM ensemble model: A regional investigation of Syria. *Engineering Applications of Computational Fluid Mechanics*, *14*(1), 90–106. https://doi.org/10.1080/19942060.2019.1683076

Huth, R. (2004). Sensitivity of local daily temperature change estimates to the selection of downscaling models and predictors. *Journal of Climate*, *17*(3), 640–652. https://doi.org/10.1175/1520-0442(2004)017%3C0640:SOLDTC%3E2.0.CO;2

IPCC. (2014). *Climate change 2013—The physical science basis: Working group I contribution to the fifth assessment report of the intergovernmental panel on climate change*. Cambridge University Press. https://doi.org/10.1017/CBO9781107415324

Kharin, V. V., & Zwiers, F. W. (2003). On the ROC score of probability forecasts. *Journal of Climate*, *16*(24), 4145–4150. https://doi.org/10.1175/1520-0442(2003)016%3C4145:OTRSOP%3E2.0.CO;2

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Y. Bengio & Y. LeCun (Eds.), *ICLR (poster)*. Retrieved from http://dblp.uni-trier.de/db/conf/iclr/iclr2015.html#KingmaB14

Klein Tank, A. M. G., Wijngaard, J. B., Können, G. P., Böhm, R., Demarée, G., Gocheva, A., et al. (2002). Daily dataset of 20th-century surface air temperature and precipitation series for the European climate assessment. *International Journal of Climatology*, *22*(12), 1441–1453. https://doi.org/10.1002/joc.773

Laprise, R., de Elía, R., Caya, D., Biner, S., Lucas-Picher, P., Diaconescu, E., et al. (2008). Challenging some tenets of regional climate modelling. *Meteorology and Atmospheric Physics*, *100*(1), 3–22. https://doi.org/10.1007/s00703-008-0292-9

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. https://doi.org/10.1109/5.726791

Lee, R. (2015). Storm track biases and changes in a warming climate from an extratropical cyclone perspective using CMIP5 PhD thesis. University of Reading. Retrieved from http://centaur.reading.ac.uk/79416

Legasa, M. N., Manzanas, R., Calviño, A., & Gutiérrez, J. M. (2022). A posteriori random forests for stochastic downscaling of precipitation by predicting probability distributions. *Water Resources Research*, *58*(4), e2021WR030272. https://doi.org/10.1029/2021WR030272

Lorenz, E. N. (1969). Atmospheric predictability as revealed by naturally occurring analogues. *Journal of the Atmospheric Sciences*, *26*(4), 636–646. https://doi.org/10.1175/1520-0469(1969)26%3C636:APARBN%3E2.0.CO;2

Manzanas, R., Brands, S., San-Martín, D., Lucero, A., Limbo, C., & Gutiérrez, J. M. (2015). Statistical downscaling in the tropics can be sensitive to reanalysis choice: A case study for precipitation in the Philippines. *Journal of Climate*, *28*(10), 4171–4184. https://doi.org/10.1175/JCLI-D-14-00331.1

Manzanas, R., Fiwa, L., Vanya, C., Kanamaru, H., & Gutiérrez, J. M. (2020). Statistical downscaling or bias adjustment? A case study involving implausible climate change projections of precipitation in Malawi. *Climatic Change*, *162*(3), 1437–1453. https://doi.org/10.1007/s10584-020-02867-3

Manzanas, R., Gutiérrez, J. M., Bhend, J., Hemri, S., Doblas-Reyes, F. J., Penabad, E., & Brookshaw, A. (2020). Statistical adjustment, calibration and downscaling of seasonal forecasts: A case-study for southeast Asia. *Climate Dynamics*, *54*(5), 2869–2882. https://doi.org/10.1007/s00382-020-05145-1

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

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, *7*(35), 983–999. Retrieved from http://jmlr.org/papers/v7/meinshausen06a.html

Nikulin, G., Asharaf, S., Magariño, M. E., Calmanti, S., Cardoso, R. M., Bhend, J., et al. (2018). Dynamical and statistical downscaling of a global seasonal hindcast in eastern Africa. *Climate Services*, *9*, 72–85. https://doi.org/10.1016/j.cliser.2017.11.003

Pham, Q. B., Yang, T.-C., Kuo, C.-M., Tseng, H.-W., & Yu, P.-S. (2019). Combing random forest and least square support vector regression for improving extreme rainfall downscaling. *Water*, *11*(3), 451. https://doi.org/10.3390/w11030451

Pour, S. H., Shahid, S., Chung, E.-S., & Wang, X.-J. (2018). Model output statistics downscaling using support vector machine for the projection of spatial and temporal changes in rainfall of Bangladesh. *Atmospheric Research*, *213*, 149–162. https://doi.org/10.1016/j.atmosres.2018.06.006

R Core Team. (2020). R: A language and environment for statistical computing. [Computer software manual]. Retrieved from https://www.R-project.org/

Riahi, K., Rao, S., Krey, V., Cho, C., Chirkov, V., Fischer, G., et al. (2011). RCP 8.5—A scenario of comparatively high greenhouse gas emissions. *Climatic Change*, *109*(1), 33–57. https://doi.org/10.1007/s10584-011-0149-y

Sa'adi, Z., Shahid, S., Pour, S. H., Ahmed, K., Chung, E.-S., & Yaseen, Z. M. (2020). Multi-variable model output statistics downscaling for the projection of spatio-temporal changes in rainfall of Borneo Island. *Journal of Hydro-Environment Research*, *31*, 62–75. https://doi.org/10.1016/j.jher.2020.05.002

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

Siabi, E. K., Kabobah, A. T., Akpoti, K., Anornu, G. K., Amo-Boateng, M., & Nyantakyi, E. K. (2021). Statistical downscaling of global circulation models to assess future climate changes in the Black Volta basin of Ghana. *Environmental Challenges*, *5*, 100299. https://doi.org/10.1016/j.envc.2021.100299

Themeßl, M. J., Gobiet, A., & Leuprecht, A. (2011). Empirical-statistical downscaling and error correction of daily precipitation from regional climate models. *International Journal of Climatology*, *31*(10), 1530–1544. https://doi.org/10.1002/joc.2168

Vaittinada Ayar, P., Vrac, M., Bastin, S., Carreau, J., Déqué, M., & Gallardo, C. (2016). Intercomparison of statistical and dynamical downscaling models under the EURO- and MED-CORDEX initiative framework: Present climate evaluations. *Climate Dynamics*, *46*(3), 1301–1329. https://doi.org/10.1007/s00382-015-2647-5

van Vuuren, D. P., Edmonds, J., Kainuma, M., Riahi, K., Thomson, A., Hibbard, K., et al. (2011). The representative concentration pathways: An overview. *Climatic Change*, *109*(1), 5–31. https://doi.org/10.1007/s10584-011-0148-z

von Storch, H., Zorita, E., & Cubasch, U. (1993). Downscaling of global climate change estimates to regional scales: An application to Iberian rainfall in wintertime. *Journal of Climate*, *6*(6), 1161–1171. https://doi.org/10.1175/1520-0442(1993)006<1161:DOGCCE>2.0.CO;2

Vrac, M., & Friederichs, P. (2015). Multivariate—Intervariable, spatial, and temporal—Bias correction. *Journal of Climate*, *28*(1), 218–237. https://doi.org/10.1175/JCLI-D-14-00059.1

Wilby, R., Dawson, C., & Barrow, E. (2002). SDSM—A decision support tool for the assessment of regional climate change impacts. *Environmental Modelling & Software*, *17*(2), 145–157. https://doi.org/10.1016/S1364-8152(01)00060-3

Wootten, A. M., Dixon, K. W., Adams-Smith, D. J., & McPherson, R. A. (2020). Statistically downscaled precipitation sensitivity to gridded observation data and downscaling technique. *International Journal of Climatology*, *41*(2), 980–1001. https://doi.org/10.1002/joc.6716

Xu, R., Chen, N., Chen, Y., & Chen, Z. (2020). Downscaling and projection of multi-CMIP5 precipitation using machine learning methods in the Upper Han River Basin. *Advances in Meteorology*, *2020*, 8680436. https://doi.org/10.1155/2020/8680436

Zorita, E., & von Storch, H. (1999). The analog method as a simple statistical downscaling technique: Comparison with more complicated methods. *Journal of Climate*, *12*(8), 2474–2489. https://doi.org/10.1175/1520-0442(1999)012%3C2474:TAMAAS%3E2.0.CO;2