







On the limitations of deep learning for statistical downscaling of climate change projections: The transferability and the extrapolation issues

Alfonso Hernanz¹  | Carlos Correa¹  | Juan-Carlos Sánchez-Perrino¹  |
Ignacio Prieto-Rico²  | Esteban Rodríguez-Guisado¹  | Marta Domínguez¹  |
Ernesto Rodríguez-Camino¹ 

¹Spanish Meteorological Agency (AEMET), Madrid, Spain

²AEMET, A Coruña, Spain

Correspondence

Alfonso Hernanz, Spanish Meteorological Agency (AEMET), Madrid 28040, Spain.

Email: ahernanzl@aemet.es

Abstract

Convolutional neural networks (CNNs) have become one of the state-of-the-art techniques for downscaling climate projections. They are being applied under Perfect-Prognosis (trained in a historical period with observations) and hybrid approaches (as Regional Climate Models (RCMs) emulators), with satisfactory results. Nevertheless, two important aspects have not been, to our knowledge, properly assessed yet: (1) their performance as emulators for other Earth System Models (ESMs) different to the one used for training, and (2) their performance under extrapolation, that is, when applied outside of their calibration range. In this study, we use UNET, a popular CNN, to assess these two aspects through two pseudo-reality experiments, and we compare it with simpler emulators: an interpolation and a linear regression. The RCA4 regional model, with 0.11° resolution over a complex domain centered in the Pyrenees, and driven by the CNRM-CM5 global model is used to train the emulators. Two frameworks are followed for the training: predictors are taken (1) from the upscaled RCM and (2) from the ESM. In both frameworks, the performance of the UNET when applied for other ESMs different to the one used for training is considerably worse, indicating poor generalization. For the linear method a similar deterioration is seen, so this limitation does not seem method specific but inherent to the task. For the second experiment, the emulators are trained in present and evaluated in future, under extrapolation. While averaged aspects such as the mean values are well simulated in future, significant biases (up to 5°C) appear when assessing warm extremes. These biases are larger by UNET than those produced by the linear method. This limitation suggests that, for variables such as temperature, with a marked signal of change and a strong linear relationship with predictors, simple linear methods might be more appropriate than the sophisticated deep learning techniques.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *Atmospheric Science Letters* published by John Wiley & Sons Ltd on behalf of Royal Meteorological Society.

KEYWORDS

convolutional neural networks, deep learning, emulators, EURO-CORDEX, evaluation, extrapolation, pseudo-reality, regional climate models, statistical downscaling

1 | INTRODUCTION

There is a growing need for high resolution climate change projections for impact and adaptation studies. This need is usually met by increasing the resolution of global simulations with some sort of downscaling. Two main approaches are possible: dynamical and statistical downscaling (SD), and they have been widely reviewed (Benestad et al., 2008; Charles et al., 2004; Huth et al., 2015; Jacob et al., 2020; Maraun et al., 2010; Rummukainen, 2010; Trzaska & Schnarr, 2014; Wilby et al., 2004; Wilby & Wigley, 1997; Zorita & von Storch, 1999). Dynamical downscaling usually consists in nesting a high-resolution model, such as a Regional Climate Model (RCM), within a lower-resolution model, such as an Earth System Model (ESM), while SD is based on the existence of statistical relationships between large scale variables (predictors) and local weather (predictands). Since dynamical downscaling is based on physical laws, one of its advantages is the physical consistency among downscaled variables. Nonetheless, its computational expense makes it difficult to use this strategy for the generation of large ensembles (Trzaska & Schnarr, 2014). On the other hand, SD is less computationally expensive, allowing exploration of uncertainties through the generation of large ensembles, but two major drawbacks are the need for historical observations and the stationarity assumption it relies on; SD is based on the assumption that the predictors/predictand relationships are maintained under future climate change, which is not granted and cannot be directly tested due to the lack of observations for the future (Charles et al., 2004; Trzaska & Schnarr, 2014; Wilby et al., 2004).

Several strategies have been proposed to indirectly assess the transferability of SD methods to different climates though. One possible approach is to use the coldest/wettest years of a historical record to train methods, and then evaluate them over the warmest/driest years (see Gutiérrez et al., 2013; Olmo & Bettolli, 2022; San-Martín et al., 2017) for temperature/precipitation. This approach is limited to the observed variability though. Another approach is to downscale future simulations and to study the impact of downscaling on the long term trends. Ideally, downscaling techniques should preserve ESMs trends in the large scale (see Baño-Medina et al., 2021; Hernanz et al., 2023; Vandal et al., 2019; Xu et al., 2020). This approach is limited to the analysis of

the spatial scales in which ESMs operate (coarse resolution), and might hide imperfections both in the spatial and temporal finer scales. And a third approach is to use pseudo-observations (RCM outputs) to train and test (in the present and future, respectively) statistical methods (see Charles et al., 1999; Gaitan et al., 2014; Hernanz et al., 2022a). This strategy allows to detect errors in the finer scales and to explore a wider range of climate change than the first approach, but the use of pseudo-observations instead of actual observations introduces an additional source of uncertainty.

Recently a new hybrid approach, RCM emulators (Doury et al., 2023), has been proposed combining the advantages of both dynamical and statistical downscaling. This strategy makes use of statistical methods to emulate the behavior of an RCM. Thus, while traditionally SD methods are trained with observations in a historical period, emulators are trained using the RCM outputs as predictands, so their training is not restricted to the historical climate. This approach presents several advantages over traditional SD and RCMs: (1) the use of future simulations for the training enables emulators to be trained with a wider range of climate states than SD under Perfect Prognosis (where the training is done in a historical period with observations), avoiding thus potential problems arising from the stationarity assumption. And (2), with the use of emulators, large ensembles can be produced from a reduced set of RCM simulations. A historical scenario and a high end emission scenario can be used to train an emulator and then produce future intermediate scenarios at low computational expense. On the other hand, the main disadvantage of this approach is that RCMs biases are maintained and must be adjusted or taken into account.

Deep learning (DL; see LeCun et al., 2015; Schmidhuber, 2015, for an overview) is a growing field with many applications, including climate downscaling. Convolutional Neural Networks (CNNs) can deal with large amounts of data and they present an important advantage over other statistical methods; their ability to extract high-level spatial features automatically (LeCun et al., 1998; LeCun & Bengio, 1995). CNNs have become one of the main *state-of-the-art* downscaling techniques, both as Perfect Prognosis SD methods and as hybrid approaches (see Baño-Medina et al., 2020, 2021; Höhlein et al., 2020; Liu et al., 2023; Passarella et al., 2022; Serifi et al., 2021; Vandal et al., 2017, 2019). The particular implementation

UNET (Ronneberger et al., 2015) has been widely used for image recognition with great performance and different variations have been also satisfactorily applied to climate downscaling (Doury et al., 2023; Sha et al., 2020a, 2020b; Sharma & Mitra, 2022).

Doury et al. (2023) proposed an emulator based on UNET to reproduce an RCM behavior. First, the emulator was trained with a RCM nested on an ESM both in a historical scenario and in the Representative Concentration Pathway (RCP) 8.5 (see IPCC, 2013), and then the emulator was evaluated over an intermediate scenario (RCP4.5), driven by the same ESM, with good results. Wang et al. (2021) used a similar emulator for precipitation, also with satisfactory results. Nevertheless, being the purpose of emulators to produce large ensembles (multiple scenarios and models) at low computational cost, their performance over ESMs different to the one used for calibration should be assessed.

Additionally, to our knowledge, the extrapolation capability of CNNs has not yet been assessed using pseudo-observations, so the finer scales can be analyzed. Hsieh (2009) and Hernanz et al. (2022b) pointed to important potential problems of neural networks and other machine learning algorithms when they are applied beyond their calibration range.

In this study, we expand the emulator proposed by Doury et al. (2023) to assess its performance driven by other ESMs different to the one used for training, and we also analyze the behavior of the deep learning tool UNET under extrapolation. The document is organized as follows. First, in Section 2 a description of the datasets, experiments, methods and evaluation metrics is provided. In Section 3 evaluation results are shown and commented on. And finally, discussion and conclusions are presented in Section 4.

2 | METHODOLOGY

In the following subsections, a description of the datasets used, the experiments design, the emulator architecture and the evaluation metrics is provided.

2.1 | Data and experiments design

This study focuses on the downscaling of surface daily mean temperature over a small but complex domain centered over the Pyrenees, including part of the Mediterranean and Atlantic coasts of Spain and France, and the Balearic Islands (see Figure 1). The predictand consists of 2345 land grid points, from a 64×64 RCM grid points domain, over a rotated grid of 0.11° . Predictors cover a

larger region (55.5° N, -9° W, 33° N, 13.5° E) corresponding to a 16×16 ESM grid points domain with a resolution of 1.5° (all ESMs are interpolated to the same grid using a bilinear interpolation). For this study, the following datasets have been used (see Table 1).

The RCA4 model (Samuelsson et al., 2011), participant in EURO-CORDEX (Jacob et al., 2014), is the RCM to be emulated with statistical methods. RCA4 is driven by four different ESMs participants in the Coupled Model Intercomparison Project Phase 5 (CMIP5, Taylor et al., 2012): CNRM-CM5 (Volodire et al., 2013), IPSL-CM5A-MR (Dufresne et al., 2013), HadGEM2-ES (Martin et al., 2011) and NorESM1-M (Bentsen et al., 2012; Iversen et al., 2013), all of them corresponding to the r1i1p1 run.

The first experiment tests the generalization of the emulator to different ESMs. In this experiment the emulator is trained with RCA4 driven by CNRM-CM5 under RCP8.5, and then it is applied and evaluated for the four ESMs under the intermediate scenario RCP4.5 (2006–2100). The reason for this choice is that the final purpose of RCM emulators is to replace some RCM simulations by the emulator. The best strategy for this is to produce an extreme scenario with the RCM and then generate intermediate scenarios (avoiding thus potential problems related with extrapolation) with the emulator. The relationship between large scale and local variables for a RCM is stronger if the large scale variables are taken from the RCM itself and not from the driving ESM (see Doury et al., 2023). Thus, an additional set of predictors is used in some cases, and it consists in the RCM upscaled to the coarse resolution (1.5°) using a conservative interpolation, what is referred to as Upscaled Regional Climate Model (UPRCM). Two evaluation frameworks are explored: (1) the Perfect Model Framework and (2) the Model World Framework (see Doury et al., 2023). In the Perfect Model Framework, predictors for training are taken from the UPRCM, and in the Model World Framework they are taken from the driving ESM instead. In this study, we have applied and evaluated the emulator for the four ESMs plus the UPRCM, following both training frameworks. It should be noticed that evaluating the emulator for the UPRCM does not represent a realistic practical case, but a theoretical optimum benchmark to compare with.

The second experiment assesses the extrapolation capability of the emulator. In this experiment, the emulator is trained under the intermediate scenario RCP4.5 and only in the present (2006–2025) and evaluated under the extreme scenario RCP8.5 in the future (2081–2100). For this experiment only the UPRCM has been used (both for training and testing), aiming for the best possible conditions for the emulator. It should be noticed that the extrapolation issue is not as crucial for emulators as it

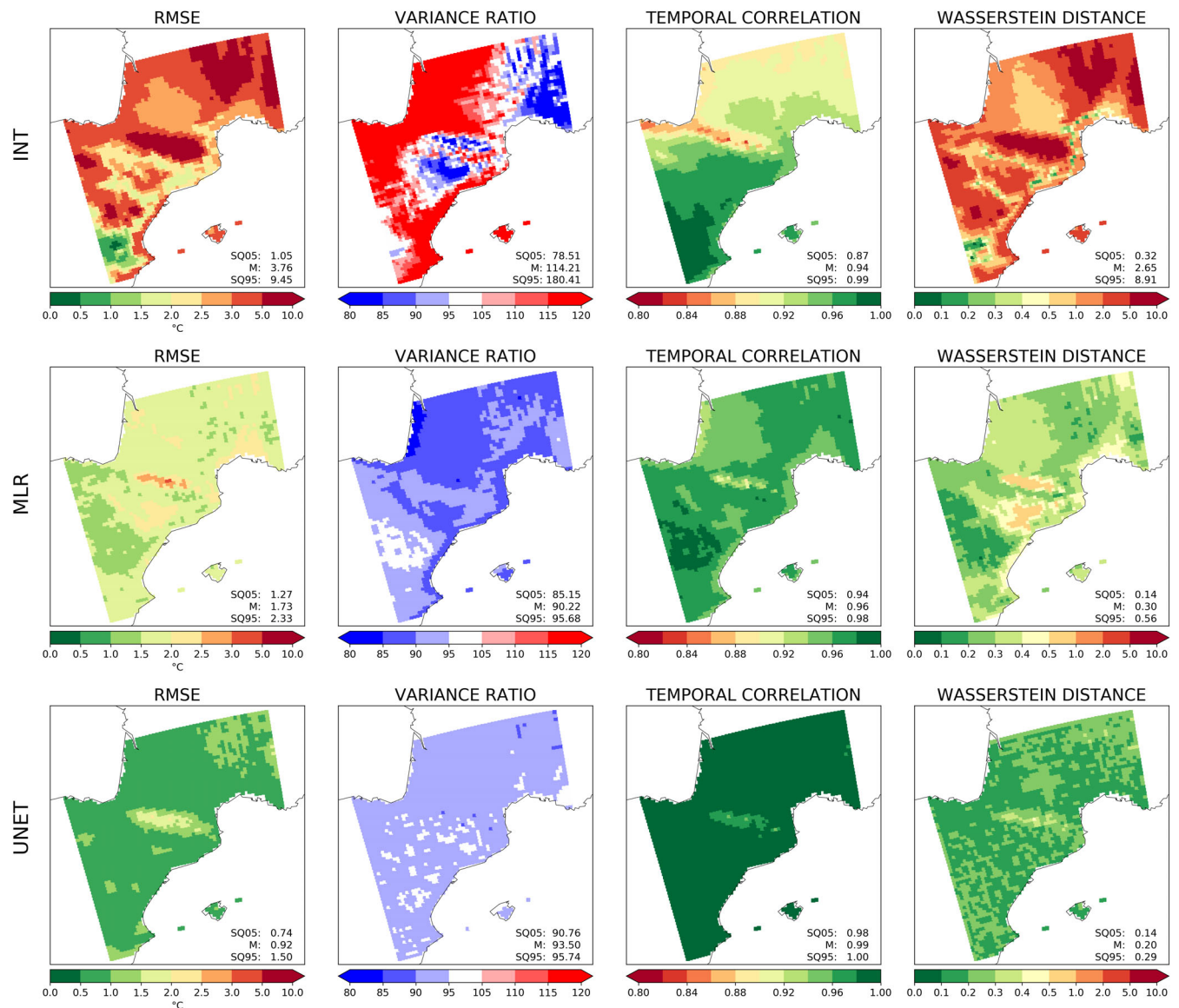


FIGURE 1 Daily RMSE (°C), variance ratio (%), temporal correlation and Wasserstein distance (in columns) by INT, MLR and UNET (in rows) in the complete period (2006–2100). The models have been trained and evaluated in the Perfect Model Framework (trained with the UPRCM driven by CNRM-CM5 under RCP8.5 and evaluated with the UPRCM driven by CNRM-CM5 under RCP4.5).

is for the Perfect Prognosis approach. Being emulators trained not only with historical data but also with future projections, the range of the training dataset is wider and extrapolation problems are not significant. Nevertheless, this experiment aims at highlighting the extrapolation problems that UNET can suffer from when applied under a Perfect Prognosis approach. The reason for assessing potential weaknesses of the Perfect Prognosis approach using the hybrid approach is that it allows to evaluate statistical methods under a larger extrapolation. The variables used as predictors are: temperature, zonal wind and meridional wind at 850 hPa and 500 hPa, geopotential height at 500 hPa, specific humidity at 850 hPa and mean sea level pressure. The choice of predictors has been

conditioned by availability, prioritizing predictors commonly used in statistical downscaling and avoiding variables strongly dependent on the model parameterizations, such as cloud cover or radiation, for example. They are standardized using their mean and standard deviation for the reference period (2006–2035) over the emission scenario used for training. Doury et al. (2023) proposed applying predictors a smoothing filter (averaging values over 3×3 grid boxes) previous to the standardization and the downscaling. This proceeding was based on Klaver et al. (2020), where the conclusion that the effective resolution of ESMs is often larger (about 3 times) than their nominal resolution was reached. Such a preprocess has been done but similar results were reached (not shown).

TABLE 1 Regional Climate Model and Earth System Models used. CNRM-CM5 is the ESM used for driving the RCA4 RCM during the training.

Model type	Model name	Institution	Resolution (lon × lat)	Reference
RCM	RCA4	Rosby Centre, Swedish Meteorological and Hydrological Institute (SMHI), Sweden	0.11° (rotated grid)	Samuelsson et al. (2011)
ESM (training)	CNRM-CM5	Centre National de Recherches Météorologiques/ Centre Européen de Recherche et Formation Avancée en Calcul Scientifique (CNRM-CERFACS), France	1.4° × 1.4°	Voldoire et al. (2013)
ESM	IPSL-CM5A-MR	Institut Pierre-Simon Laplace (IPSL), France	2.5° × 1.2°	Dufresne et al. (2013)
ESM	HadGEM2-ES	Met Office, UK	1.9° × 1.2°	Martin et al. (2011)
ESM	NorESM1-M	Norwegian Meteorological Institute (MET Norway)	1.89° × 2.50°	Bentsen et al. (2012), Iversen et al. (2013)

2.2 | Emulator description

The emulator here used is the same proposed by Doury et al. (2023) which in turn is an adaptation of the original UNET (Ronneberger et al., 2015). UNET is a popular architecture designed for biomedical purposes. It is a fully convolutional neural network, consisting of a series of four encoding blocks followed by a series of four decoding blocks which are, in addition, connected by bridges. Encoding blocks are formed by three layers: two of them convolutional (with 3×3 kernels) and then a max pooling (down-sampling operation through the use of maximum filters) with 2×2 filters. Similarly, decoding blocks are inversely formed by a 2×2 up-sampling and two 3×3 convolutional layers.

This architecture is usually represented by a U-shaped network which gives the model its name. In its original design, UNET was used for image segmentation, which is a classification problem. In this case, UNET tackles a regression problem, because temperature is a continuous numerical variable. Thus, the original UNET has been modified in two main ways: (1) the output layer uses a REctified Linear Unit function (ReLU, more appropriate for regression tasks than the original sigmoid) and (2) the loss function and metric used are the root mean squared error (instead of the original binary cross entropy and accuracy, respectively). An Adam optimizer (Kingma & Ba, 2014) with learning rate of 0.005 has been used, with 100 epochs and a batch size of 32. Overfitting has been handled by the use of early stopping (stopping the gradient descent once no more improvement is found after k iterations in a validation set, with k typically called patience) with a patience of 15. For a more detailed description, see Doury et al. (2023).

This emulator based on UNET is compared to simpler benchmarks: an emulator consisting on a bilinear interpolation (INT) and another one consisting in a multiple linear regression (MLR). For the MLR, predictors are taken from the four nearest neighbors and interpolated to each target point.

2.3 | Evaluation metrics

For the evaluation of the emulators we have followed the metrics used in Doury et al. (2023). Three different aspects are evaluated:

1. Daily metrics: the Root Mean Squared Error (RMSE), the variance ratio (%), the temporal correlation (through the Pearson's correlation coefficient) and the 1-D Wasserstein distance. The Wasserstein distance is a metric based on the optimal transport theory (Villani, 2009) and it measures the similarity of the statistical distributions.
2. Present climatology: mean temperature and 99th percentile in 2006–2025.
3. Climate change signal: the delta changes in the mean temperature and in the 99th percentile between the future period (2081–2100) and the present period (2006–2025).

These metrics are calculated at grid point scale. When presented in the form of maps, they are accompanied by their means (M) and their super-quantiles of order 0.05 (SQ05) and 0.95 (SQ95) in order to summarize the maps in a few values. The super-quantile α is defined as the mean of all the values larger (resp. smaller) than the quantile of order α .

3 | RESULTS

3.1 | First experiment: RCM emulator generalization

3.1.1 | Perfect model framework

In the first experiment, the emulator is trained under RCP8.5 and evaluated in RCP4.5. First, the strengths of the emulator are to be proven. Figures 1 and 2 show scores for the emulator UNET compared to the interpolation and the linear method in the Perfect Model Approach (trained and evaluated for UPRCM). In this framework, daily scores by UNET are systematically better than by the other methods (Figure 1). RMSEs by the UNET (mean value of 0.92°C) are considerably lower than those for INT and MLR (3.76°C and 1.73°C , respectively). The variance ratios by UNET (mean value of 93.50%) are closer to 100% than by INT and MLR (114.21% and 90.22%, respectively). Correlations by UNET (mean value of 0.99) are higher than for INT and MLR (0.94 and 0.96, respectively). Wasserstein distances by UNET (mean value of 0.20) are lower than those by INT and MLR (2.65 and 0.30, respectively). As for the climatology and delta change (Figure 2), UNET also improves the other two methods, although for the mean temperature MLR achieves similar scores. For the mean temperature in the present period (2006–2025), UNET and MLR present low biases (mean values of 0.13°C and -0.03°C , respectively) while INT presents larger biases (mean values of 1.83°C). The delta change in the mean temperature, though, is well captured by all methods, with mean biases lower than 0.05°C in absolute value. It is in the 99th percentile when the improvements of UNET over MLR become more apparent. For the present climatology, UNET presents very low biases (mean value of -0.47°C), considerably lower than by INT and MLR (1.67°C and -1.46°C , respectively). And for the delta changes in the 99th percentile, biases by UNET (mean value of -0.01°C) are also clearly lower than by the other two methods (0.54°C and 0.37°C for INT and MLR, respectively). Only the mean values have been commented, but the extremes of the spatial distributions (SQ05 and SQ95) lead to the same conclusions

3.1.2 | Model world framework

Now that the good performance of UNET under the Perfect Model Framework has been demonstrated, let us analyze its performance when applied for other ESMs. RMSEs by UNET (mostly below 1°C) are considerably lower than by MLR (around $1.5\text{--}2^{\circ}\text{C}$) for the UPRCM.

But when UNET is applied for the other ESMs (even for CNRM-CM5, the one used for driving the RCM in the training dataset), its RMSEs increase up to around $2\text{--}2.5^{\circ}\text{C}$, being similar to the ones by MLR (Figure 3). Thus, it has been proved that despite the good performance of UNET when applied for the same ESM that it has been trained with (UPRCM in this case, but similar results are reached with CNRM-CM5, not shown), when applied for other ESMs its performance is considerably worse

3.2 | Second experiment: Extrapolation

For the second experiment, the one focusing on extrapolation, we present results for the present and future climatologies as well as for the delta changes for the mean temperature and for the 99th percentile. Figure 4 shows how the mean climatology and delta change in it is well captured by UNET and MLR, but as it has been mentioned, the analysis only of averaged aspects (either temporary or spatially) can hide imperfections in the finer scales. When analyzing the 99th percentile, UNET presents low biases (around -1°C to -0.5°C) in the present climatology (inside its calibration range), but for the future climatology, these extreme values present significant biases (mostly between -0.5°C and 2°C , but up to 5°C in some cases), larger than those given by the linear method (as well as for the delta change as a consequence).

4 | DISCUSSION AND CONCLUSIONS

Deep learning methods based on CNNs are *state-of-the-art* downscaling methods, being used both for statistical downscaling under Perfect Prognosis and for hybrid approaches, such as RCM emulators. In this study, we have identified two important limitations for both applications: (1) RCM emulators appear to be ESM dependent (i.e., their performance is considerably different when evaluated for ESMs different to the one used during their calibration) and (2) their performance under extrapolation (values outside of the calibration range) is poor. For this study we have used the popular UNET, a specific implementation that has shown great performance in many fields including climate downscaling.

The first limitation is not exclusive of CNNs, but it is also seen in linear methods. That indicates that the predictors/predictand relationships established by the RCM are different for different driving ESMs. A possible explanation for this finding could be related with some

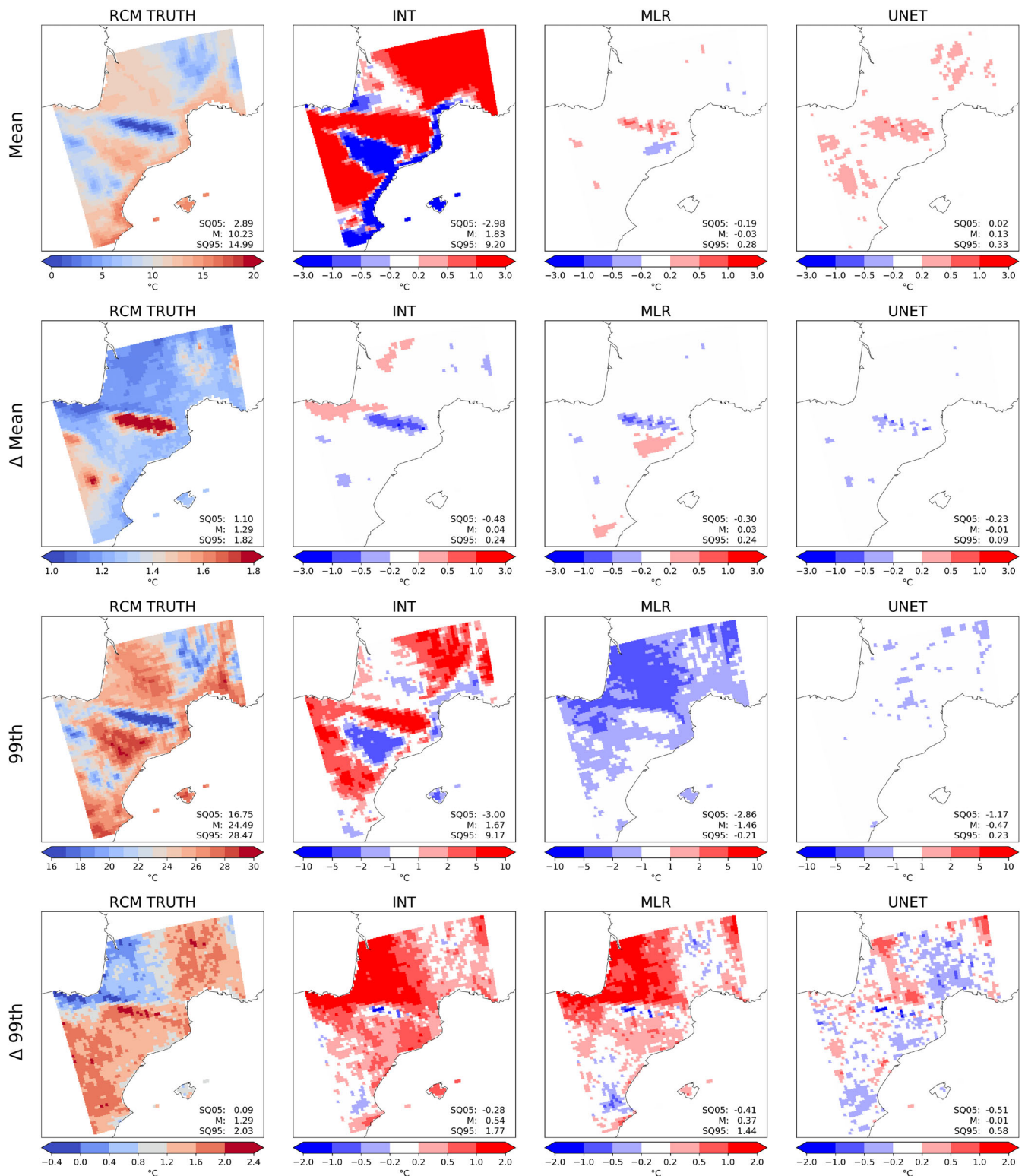


FIGURE 2 RCM truth and bias by INT, MLR and UNET (in columns) for the mean temperature (°C, present climatology in first row and future delta change in second row) and for the 99th percentile (°C, present climatology in third row and future delta change in fourth row). The present climatology corresponds to 2006–2025 and the future delta change corresponds to the difference between 2081–2100 and 2006–2025. The models have been trained and evaluated in the Perfect Model Framework (trained with the UPRCM driven by CNRM-CM5 under RCP8.5 and evaluated with the UPRCM driven by CNRM-CM5 under RCP4.5).

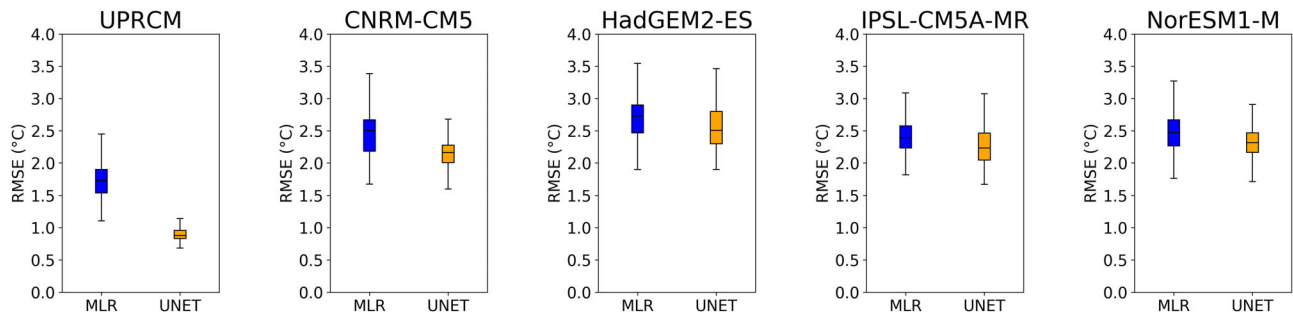


FIGURE 3 Daily RMSE ($^{\circ}\text{C}$) for the complete period (2006–2100). The models have been trained in the Perfect Model Framework (UPRCM driven by CNRM-CM5 under RCP8.5) and evaluated over the UPRCM driven by CNRM-CM5, CNRM-CM5, HadGEM2-ES, IPSL-CM5A-MR and NorESM1-M (from left to right) under RCP4.5. Each box (MLR in blue and UNET in orange) summarizes the distribution of the 2345 grid points by the median and the quartiles; whiskers extend to a maximum of 1.5 times the interquartile range.

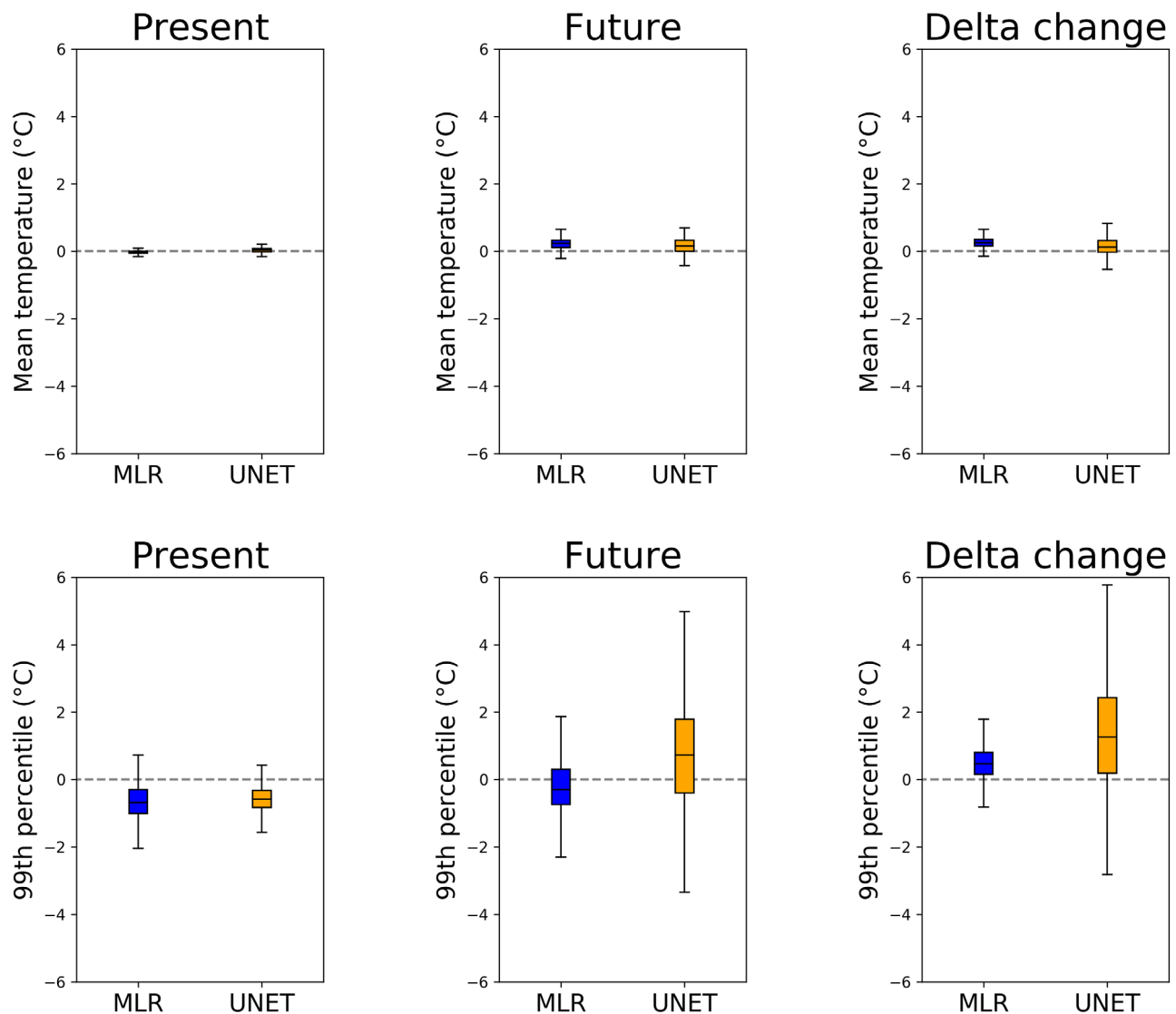


FIGURE 4 Bias in the mean temperature ($^{\circ}\text{C}$, first row) and the 99th percentile ($^{\circ}\text{C}$, second row) in the present climatology (first column, 2006–2025), the future climatology (second column, 2081–2100) and the delta change (third column, difference between 2081–2100 and 2006–2025). The models have been trained and evaluated in the Perfect Model Framework (UPRCM driven by CNRM-CM5 under RCP4.5 for training and RCP8.5 for evaluation). Each box (MLR in blue, and UNET in orange) summarizes the distribution of the 2345 grid points by the median and the quartiles; whiskers extend to a maximum of 1.5 times the interquartile range.

overfitting by the emulators. Another possible explanation could be that the set of predictors used is not enough to explain all the predictand variability, and predictors containing important information might have been missing (e.g., aerosols, clouds, radiation, surface processes, or other variables dependent on each ESM parameterizations). In particular, near surface temperature is also highly dependent on soil water content which in turn is responsible for the partition of sensible and latent surface heat fluxes affecting near surface temperature and humidity. Both heat fluxes are the main mechanism to turn back energy into the atmosphere from the land surface. Consequently, the Bowen ratio (the ratio of sensible to latent heat flux), not usually contemplated among the atmospheric predictors, would help to further explain the temperature variability (Rodríguez-Camino & Avissar, 1998). Further investigation in this way might be fruitful in order to build, if possible, emulators capable of generalizing for different ESMs. These results point to only a moderate potential benefit on the use of RCM emulators. Being a large ensemble composed of multiple emission scenarios (N) and multiple driving ESMs (M), for a single RCM, the scenario/ESM matrix does not seem feasible to be filled with emulators from only a few RCM simulations. Had results been similar for other ESMs, only 2 RCMs simulations (a historical one and a high-end one) would be enough to fill the $N \times M$ matrix. This way $2 \times M$ RCMs simulations are needed (historical + high-end, for each driving ESM) for emulators to produce accurate results.

As for the second limitation, being extrapolation a well-known potential issue for any statistical method, errors under extrapolation in the fine spatial and temporal scales are not often assessed when evaluating new methods. In this study, we have demonstrated how CNNs trained in the present can reproduce accurately the mean climatology both in present and under climate change (averaged aspects), but for extreme temperatures (the tail of the distribution), important errors emerge in the future climate (outside of the calibration range). This is rarely assessed, and such errors can often remain hidden if only averaged aspects are evaluated. Nonetheless, these errors can lead to wrong conclusions for impact and adaptation studies. For variables such as temperature, where the predictors/predictand relationships are quite linear and the signal of change is strong (large amount of data projected outside of the calibration range), simple linear methods might be more suitable than sophisticated deep learning techniques.

Needless to say that these experiments have been carried out for a particular RCM and configuration, but an expansion to other RCMs/configurations would lead to more robust conclusions. Similarly, these conclusions have been reached using a particular deep learning

approach and a set of evaluation metrics, but other are possible and might lead to different conclusions.

AUTHOR CONTRIBUTIONS

Alfonso Hernanz: Conceptualization; data curation; visualization; writing – original draft. **Carlos Correa Guinea:** Writing – review and editing. **Juan Carlos Sánchez Perrino:** Writing – review and editing. **Ignacio Prieto:** Writing – review and editing. **Esteban Rodríguez-Guisado:** Writing – review and editing. **Marta Domínguez:** Writing – review and editing. **Ernesto Rodríguez-Camino:** Writing – review and editing.

ACKNOWLEDGEMENTS

We would like to acknowledge the World Climate Research Programme's Working Group on Regional Climate, and the Working Group on Coupled Modelling, former coordinating body of CORDEX and responsible panel for CMIP5. We also thank the Climate Modelling Groups (listed in Table 1 of this paper) for producing and making available their model output. We also acknowledge the Earth SystemGrid Federation infrastructure, an international effort led by the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison, the European Network for Earth System Modelling and other partners in the Global Organisation for Earth System Science Portals (GO-ESSP). Finally, we would like to thank two anonymous reviewers for their very constructive comments.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflict of interest.

DATA AVAILABILITY STATEMENT

All datasets here used are freely available at the Earth System Grid Federation nodes (<https://esgf-node.llnl.gov/search/cmip5/> for CMIP5 ESMs and <https://esgf-data.dkrz.de/search/cordex-dkrz/> for CORDEX RCMs). The emulator is and adaptation of <https://github.com/antoinedoury/RCM-Emulator>.

ORCID

Alfonso Hernanz  <https://orcid.org/0000-0003-1091-0422>

Carlos Correa  <https://orcid.org/0000-0003-3049-6185>

Juan-Carlos Sánchez-Perrino  <https://orcid.org/0000-0003-4916-6709>

Ignacio Prieto-Rico  <https://orcid.org/0009-0007-7883-4936>

Esteban Rodríguez-Guisado  <https://orcid.org/0000-0002-9608-1751>

Marta Domínguez  <https://orcid.org/0000-0001-7840-5516>

Ernesto Rodríguez-Camino  <https://orcid.org/0000-0002-1565-2373>

REFERENCES

- Baño-Medina, J., Manzanar, R. & Gutiérrez, J.M. (2021) On the suitability of deep convolutional neural networks for continental-wide downscaling of climate change projections. *Climate Dynamics*, 57, 2941–2951. Available from: <https://doi.org/10.1007/s00382-021-05847-0>
- Baño-Medina, J., Manzanar, R., Gutierrez, J.M. & Gutiérrez, J.M. (2020) Configuration and intercomparison of deep learning neural models for statistical downscaling. *Geoscientific Model Development*, 13(4), 2109–2124. Available from: <https://doi.org/10.5194/gmd-13-2109-2020>
- Benestad, R.E., Chen, D. & Hanssen-Bauer, I. (2008) *Empirical-statistical downscaling*. Singapore: World Scientific Publishing Company. Available from: <https://doi.org/10.1142/6908>
- Bentsen, M., Bethke, I., Debernard, J.B., Iversen, T., Kirkevåg, A., Seland, Ø. et al. (2012) The Norwegian Earth System Model, NorESM1-M – part 1: description and basic evaluation. *Geoscientific Model Development Discussions*, 5, 2843–2931. Available from: <https://doi.org/10.5194/gmdd-5-2843-2012>
- Charles, S., Bates, B., Whetton, P. & Hughes, J. (1999) Validation of downscaling models for changed climate conditions: case study of southwestern Australia. *Climate Research*, 12, 1–14. Available from: <https://doi.org/10.3354/cr012001>
- Charles, S.P., Bates, B.C., Smith, I.N. & Hughes, J.P. (2004) Statistical downscaling of daily precipitation from observed and modelled atmospheric fields. *Hydrological Processes*, 18, 1373–1394. Available from: <https://doi.org/10.1002/hyp.1418> Special Issue: Scale and Scaling in Hydrology.
- Doury, A., Somot, S., Gadat, S., Ribes, A. & Corre, L. (2023) Regional climate model emulator based on deep learning: concept and first evaluation of a novel hybrid downscaling approach. *Climate Dynamics*, 60, 1751–1779. Available from: <https://doi.org/10.1007/s00382-022-06343-9>
- Dufresne, J.L., Foujols, M.A., Denvil, S., Caubel, A., Marti, O., Aumont, O. et al. (2013) Climate change projections using the IPSL-CM5 earth system model: from CMIP3 to CMIP5. *Climate Dynamics*, 40, 2123–2165. Available from: <https://doi.org/10.1007/s00382-012-1636-1>
- Gaitan, C., Hsieh, W. & Cannon, A. (2014) Comparison of statistically downscaled precipitation in terms of future climate indices and daily variability for southern Ontario and Quebec, Canada. *Climate Dynamics*, 43, 1–17. Available from: <https://doi.org/10.1007/s00382-014-2098-4>
- Gutiérrez, J.M., San-Martín, D., Brands, S., Manzanar, R. & Herrera, S. (2013) Reassessing statistical downscaling techniques for their robust application under climate change conditions. *Journal of Climate*, 26(1), 171–188. Available from: <https://doi.org/10.1175/JCLI-D-11-00687.1>
- Hernanz, A., Correa, C., Domínguez, M., Rodríguez-Guisado, E. & Rodríguez-Camino, E. (2023) Statistical downscaling in the tropics and midlatitudes: a comparative assessment over two representative regions. *Journal of Applied Meteorology and Climatology*, 62, 737–753. Available from: <https://doi.org/10.1175/JAMC-D-22-0164.1>
- Hernanz, A., García-Valero, J.A., Domínguez, M. & Rodríguez-Camino, E. (2022a) 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. Available from: <https://doi.org/10.1002/joc.7464>
- Hernanz, A., García-Valero, J.A., Domínguez, M. & Rodríguez-Camino, E. (2022b) A critical view on the suitability of machine learning techniques to downscale climate change projections: illustration for temperature with a toy experiment. *Atmospheric Science Letters*, 23(6), e1087. Available from: <https://doi.org/10.1002/asl.1087>
- Höhlein, K., Kern, M., Hewson, T. & Westermann, R. (2020) A comparative study of convolutional neural network models for wind field downscaling. *Meteorological Applications*, 27, e1961. Available from: <https://doi.org/10.1002/met.1961>
- Hsieh, W. (2009) *Machine learning methods in the environmental sciences: neural networks and kernels*. Cambridge: Cambridge University Press. Available from: <https://doi.org/10.1017/CBO9780511627217>
- Huth, R., Mikšovský, J., Štěpánek, P., Belda, M., Farda, A., Chláková, Z. et al. (2015) Comparative validation of statistical and dynamical downscaling models on a dense grid in central Europe: temperature. *Theoretical and Applied Climatology*, 120, 533–553. Available from: <https://doi.org/10.1007/s00704-014-1190-3>
- IPCC. (2013) Climate change 2013: the physical science basis. In: Stocker, T.F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S.K., Boschung, J. et al. (Eds.) *Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change*. Cambridge, UK and New York, NY, USA: Cambridge University Press, p. 1535. Available from: <https://doi.org/10.1017/CBO9781107415324>
- Iversen, T., Bentsen, M., Bethke, I., Debernard, J.B., Kirkevåg, A., Seland, Ø. et al. (2013) The Norwegian earth system model, NorESM1-M – part 2: climate response and scenario projections. *Geoscientific Model Development*, 6, 389–415. Available from: <https://doi.org/10.5194/gmd-6-389-2013>
- Jacob, D., Petersen, J., Eggert, B., Alias, A., Christensen, O.B., Bouwer, L.M. et al. (2014) EURO-CORDEX: new high-resolution climate change projections for European impact research. *Regional Environmental Change*, 14(2), 563–578. Available from: <https://doi.org/10.1007/s10113-013-0499-2>
- Jacob, D., Teichmann, C., Sobolowski, S., Katragkou, E., Anders, I., Belda, M. et al. (2020) Regional climate downscaling over Europe: perspectives from the EURO-CORDEX community. *Regional Environmental Change*, 20, 1–20. Available from: <https://doi.org/10.1007/s10113-020-01606-9>
- Kingma, D.P. & Ba, J. (2014) Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Klaver, R., Haarsma, R., Vidale, P.L. & Hazeleger, W. (2020) Effective resolution in high resolution global atmospheric models for climate studies. *Atmospheric Science Letters*, 21(4), 1–8. Available from: <https://doi.org/10.1002/asl.952>
- LeCun, Y. & Bengio, Y. (1995) *Convolutional networks for images, speech, and time series*. The handbook of brain theory and neural networks: MIT Press, pp. 255–258.
- LeCun, Y., Bengio, Y. & Hinton, G. (2015) Deep learning. *Nature*, 521(7553), 436–444. Available from: <https://doi.org/10.1038/nature14539>
- LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. (1998) Gradient-based learning applied to document recognition. *Proceedings of*

- the *IEEE*, 86(11), 2278–2323. Available from: <https://doi.org/10.1109/5.726791>
- Liu, G., Powell, B. & Friedrich, T. (2023) Climate downscaling for regional models with a neural network: a Hawaiian example. *Progress in Oceanography*, 215, 103047. Available from: <https://doi.org/10.1016/j.pocean.2023.103047>
- Maraun, D., Wetterhall, F., Ireson, A.M., Chandler, R.E., Kendon, E.J., Widmann, M. et al. (2010) Precipitation downscaling under climate change: recent developments to bridge the gap between dynamical models and the end user. *Reviews of Geophysics*, 48, RG3003. Available from: <https://doi.org/10.1029/2009RG000314>
- Martin, T., Bellouin, N., Collins, W., Culverwell, I., Halloran, P., Hardiman, S. et al. (2011) The HadGEM2 family of Met Office Unified Model Climate configurations. *Geoscientific Model Development*, 4, 723–757. Available from: <https://doi.org/10.5194/gmd-4-723-2011>
- Olmo, M.E. & Bettolli, M.L. (2022) Statistical downscaling of daily precipitation over southeastern South America: assessing the performance in extreme events. *International Journal of Climatology*, 42(2), 1283–1302. Available from: <https://doi.org/10.1002/joc.7303>
- Passarella, L.S., Mahajan, S., Pal, A. & Norman, M.R. (2022) Reconstructing high resolution ESM data through a novel fast super resolution convolutional neural network (FSRCNN). *Geophysical Research Letters*, 49, e2021GL097571. Available from: <https://doi.org/10.1029/2021GL097571>
- Rodríguez-Camino, E. & Avissar, R. (1998) Comparison of three land-surface schemes with the Fourier amplitude sensitivity test (FAST). *Tellus A: Dynamic Meteorology and Oceanography*, 50(3), 313–332. Available from: <https://doi.org/10.3402/tellusa.v50i3.14529>
- Ronneberger, O., Fischer, P. & Brox, T. (2015) U-net: convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9351, 234–241. Available from: https://doi.org/10.1007/978-3-319-24574-4_28
- Rummukainen, M. (2010) State-of-the-art with regional climate models. *WIREs Climate Change*, 1(1), 82–96. Available from: <https://doi.org/10.1002/wcc.8>
- Samuelsson, P., Jones, C.G., Willén, U., Ullerstig, A., Gollvik, S., Hansson, U.L.F. et al. (2011) The Rossby Centre Regional Climate model RCA3: model description and performance. *Tellus A: Dynamic Meteorology and Oceanography*, 63(1), 4–23. Available from: <https://doi.org/10.1111/j.1600-0870.2010.00478.x>
- San-Martín, D., Manzanar, 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. Available from: <https://doi.org/10.1175/JCLI-D-16-0366.1>
- Schmidhuber, J. (2015) Deep learning in neural networks: an overview. *Neural Networks*, 61, 85–117. Available from: <https://doi.org/10.1016/j.neunet.2014.09.003>
- Serifi, A., Günther, T. & Ban, N. (2021) Spatio-temporal downscaling of climate data using convolutional and error-predicting neural networks. *Frontiers in Climate*, 3, 656479. Available from: <https://doi.org/10.3389/fclim.2021.656479>
- Sha, Y., Gagne, D.J., II, West, G. & Stull, R. (2020a) Deep-learning-based gridded downscaling of surface meteorological variables in complex terrain. Part I: daily maximum and minimum 2-m temperature. *Journal of Applied Meteorology and Climatology*, 59, 2057–2073. Available from: <https://doi.org/10.1175/JAMC-D-20-0057.1>
- Sha, Y., Gagne, D.J., II, West, G. & Stull, R. (2020b) Deep-learning-based gridded downscaling of surface meteorological variables in complex terrain. Part II: daily precipitation. *Journal of Applied Meteorology and Climatology*, 59(12), 2075–2092. Available from: <https://doi.org/10.1175/JAMC-D-20-0058.1>
- Sharma, S.C.M. & Mitra, A. (2022) ResDeepD: a residual super-resolution network for deep downscaling of daily precipitation over India. *Environmental Data Science*, 1, e19. Available from: <https://doi.org/10.1017/eds.2022.23>
- Taylor, K., Ronald, S. & Meehl, G. (2012) An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, 93, 485–498. Available from: <https://doi.org/10.1175/BAMS-D-11-00094.1>
- Trzaska, S. & Schnarr, E. (2014) *A Review of Downscaling Methods for Climate Change Projections: African and Latin American Resilience to Climate Change (ARCC)*. Available from: http://www.ciesin.org/documents/Downscaling_CLEARED_000.pdf
- Vandal, T., Kodra, E. & Ganguly, A.R. (2019) Intercomparison of machine learning methods for statistical downscaling: the case of daily and extreme precipitation. *Theoretical and Applied Climatology*, 137(1–2), 557–570. Available from: <https://doi.org/10.1007/s00704-018-2613-3>
- Vandal, T., Kodra, E., Ganguly, S., Michaelis, A., Nemani, R. & Ganguly, A.R. (2017) DeepSD: generating high resolution climate change projections through single image super-resolution. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Part F1296*. pp. 1663–1672. <https://doi.org/10.1145/3097983.3098004>
- Villani, C. (2009) *Optimal Transport: Old and New*, Vol. 338. Berlin, Heidelberg: Springer. Available from: <https://doi.org/10.1007/978-3-540-71050-9>
- Voltaire, A., Sanchez-Gomez, E., Méliá, D., Decharme, B., Cassou, C., Senesi, S. et al. (2013) The CNRM-CM5.1 global climate model: description and basic evaluation. *Climate Dynamics*, 40, 2091–2121. Available from: <https://doi.org/10.1007/s00382-011-1259-y>
- Wang, J., Liu, Z., Foster, I., Chang, W., Kettimuthu, R. & Kotamarthi, V.R. (2021) Fast and accurate learned multiresolution dynamical downscaling for precipitation. *Geoscientific Model Development*, 14, 6355–6372. Available from: <https://doi.org/10.5194/gmd-14-6355-2021>
- Wilby, R., Charles, S., Zorita, E., Timbal, B., Whetton, P. & Mearns, L. (2004) Guidelines for use of climate scenarios developed from statistical downscaling methods. In: *Supporting Material of the Intergovernmental Panel on Climate Change*. Geneva: IPCC. Available from: https://www.ipcc-data.org/guidelines/dgm_no2_v1_09_2004.pdf
- Wilby, R.L. & Wigley, T.M.L. (1997) Downscaling general circulation model output: a review of methods and limitations. *United Kingdom: Progress in Physical Geography: Earth and*

Environment, 21(4), 530–548. Available from: <https://doi.org/10.1177/030913339702100403>

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. Available from: <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, 2474–2489. Available from: [https://doi.org/10.1175/1520-0442\(1999\)012<2474:TAMAAS>2.0.CO;2](https://doi.org/10.1175/1520-0442(1999)012<2474:TAMAAS>2.0.CO;2)

How to cite this article: Hernanz, A., Correa, C., Sánchez-Perrino, J.-C., Prieto-Rico, I., Rodríguez-Guisado, E., Domínguez, M., & Rodríguez-Camino, E. (2023). On the limitations of deep learning for statistical downscaling of climate change projections: The transferability and the extrapolation issues. *Atmospheric Science Letters*, e1195. <https://doi.org/10.1002/asl.1195>