Transferability and Explainability of Deep Learning Emulators for Regional Climate Model Projections: Perspectives for Future Applications

JORGE BAÑO-MEDINA,[a] MAIALEN ITURBIDE,[a] JESÚS FERNÁNDEZ,[a] AND JOSÉ MANUEL GUTIÉRREZ[a]

[a] *Instituto de Física de Cantabria, CSIC-Universidad de Cantabria, Santander, Spain*

(Manuscript received 26 October 2023, in final form 14 June 2024, accepted 8 July 2024)

ABSTRACT: Regional climate models (RCMs) are essential tools for simulating and studying regional climate variability and change. However, their high computational cost limits the production of comprehensive ensembles of regional climate projections covering multiple scenarios and driving Global climate models (GCMs) across regions. RCM emulators based on deep learning models have recently been introduced as a cost-effective and promising alternative that requires only short RCM simulations to train the models. Therefore, evaluating their transferability to different periods, scenarios, and GCMs becomes a pivotal and complex task in which the inherent biases of both GCMs and RCMs play a significant role. Here, we focus on this problem by considering the two different emulation approaches introduced in the literature as perfect and imperfect, that we here refer to as perfect prognosis (PP) and model output statistics (MOS), respectively, following the well-established downscaling terminology. In addition to standard evaluation techniques, we expand the analysis with methods from the field of explainable artificial intelligence (XAI), to assess the physical consistency of the empirical links learnt by the models. We find that both approaches are able to emulate certain climatological properties of RCMs for different periods and scenarios (soft transferability), but the consistency of the emulation functions differs between approaches. Whereas PP learns robust and physically meaningful patterns, MOS results are GCM dependent and lack physical consistency in some cases. Both approaches face problems when transferring the emulation function to other GCMs (hard transferability), due to the existence of GCM-dependent biases. This limits their applicability to build RCM ensembles. We conclude by giving prospects for future applications.

SIGNIFICANCE STATEMENT: Regional climate model (RCM) emulators are a cost-effective emerging approach for generating comprehensive ensembles of regional climate projections. Promising results have been recently obtained using deep learning models. However, their potential to capture the regional climate dynamics and to emulate other periods, emission scenarios, or driving global climate models (GCMs) remains an open issue that affects their practical use. This study explores the potential of current emulation approaches incorporating new explainable artificial intelligence (XAI) evaluation techniques to assess the reliability and transferability of the emulators. Our findings show that the different global and regional model biases involved in the different approaches play a key role in transferability. Based on the results obtained, we provide some prospects for potential applications of these models in challenging problems.

KEYWORDS: Downscaling; Neural networks; Climate variability

1. Introduction

Regional climate models (RCMs, Giorgi 2019) are sophisticated tools widely used to produce high-resolution regional climate projections. They work by numerically solving a set of physical equations representing regional atmospheric processes and interactions with other components, such as land, over a limited continental region. RCMs are driven at their boundaries by the coarse output of a global climate model (GCM; Phillips 1956), a process often referred to as dynamical downscaling. A variety of studies (Rummukainen 2016; Soares and Cardoso 2018; Molina et al. 2022; Cardoso and Soares 2022) have assessed the added value of RCMs, pointing to a better representation of the local scale as compared to their driving GCMs. Therefore, RCM simulations constitute a valuable line of evidence in assessing the risks and adaptation strategies related to climate change at the regional scale (IPCC 2022).

The Coordinated Regional Climate Downscaling Experiment (CORDEX) coordinates the generation of regional climate projections worldwide, based on multimodel ensembles of RCM simulations spanning different sources of uncertainty, including those arising from the driving GCM or the emission scenario, among others (Jacob et al. 2020; Diez-Sierra et al. 2022). However, covering the large number of potential scenario–GCM–RCM combinations is an enormous computational challenge. As a result, the limited availability of CORDEX simulations in some regions hinders the comprehensive assessment of uncertainty in regional climate modeling (Kendon et al. 2010; Fernández et al. 2019). This has led the regional climate modeling community to look for alternatives to these costly simulations, especially as they approach the kilometer scale.

Empirical–statistical downscaling (ESD) has traditionally been a cost-effective alternative/complement to dynamical downscaling (Maraun and Widmann 2018). ESD techniques rely on observations to learn the relationship between large-scale

Corresponding author: Jorge Baño-Medina, bmedina@ifca.unican.es

DOI: 10.1175/AIES-D-23-0099.1 e230099

meteorological fields (typically from reanalysis datasets) and local surface variables of interest, such as temperature and precipitation.

Recent advances in deep learning (DL) techniques, such as deep convolutional neural networks (CNNs; LeCun and Bengio 1995), have driven the development of deep downscaling methods (see Rampal et al. 2024; Sun et al. 2024; Molina et al. 2023, for a review) and their application to downscale global climate projections over large areas (Baño-Medina et al. 2021). However, the lack of sufficient observed data available for training in many regions remains as one of the main drawbacks of ESD that limits its applicability, particularly as finer spatial resolutions are considered.

An alternative approach recently introduced to overcome this problem is RCM emulation leveraging DL techniques. These models do not rely on observations. Instead, they require an existing relatively short RCM simulation (driven by a particular GCM) to train the emulator, which learns the mapping between upper-air large-scale fields and surface target variables from the RCM. Two different RCM emulation approaches have been recently introduced in the literature, referred to as "perfect" and "imperfect" (Erlandsen et al. 2020; Boé et al. 2023; Hobeichi et al. 2023). Both use surface high-resolution RCM variables (typically temperature and/or precipitation) as targets, or predictands, but differ in the predictors used. The perfect approach uses a set of informative upscaled large-scale variables from the same RCM and can be therefore considered a hybrid implementation of the perfect prognosis (PP) downscaling approach, using observations (in this case pseudo-observations within the RCM world) for both predictors and predictand (Maraun and Widmann 2018). This approach maximizes the day-to-day correspondence and physical consistency between the input–output pairs, as both come from the same model. Here, we use well-established terminology in ESD and refer to this approach as "PP" (from perfect prognosis). On the contrary, the so-called imperfect approach uses the driving GCM fields as predictors, thus coping with the lack of perfect day-to-day correspondence and model biases in the learning process. Similarly to the previous case, using standard terminology, we refer to this approach as "MOS," in an analogy with the model output statistics (MOS) downscaling approach, which deals with model biases during learning (see, e.g., Gutiérrez et al. 2019).

Several studies have analyzed both emulation approaches independently (Doury et al. 2023; Hobeichi et al. 2023) or jointly (Boé et al. 2023; van der Meer et al. 2023), based on the comparison of particular evaluation metrics/indices between the emulated and target RCM fields. These studies show promising results to emulate an intermediate temporal period for the same GCM–RCM pair, particularly for the MOS approach, since the PP one inherits GCM–RCM biases that affect the emulated fields. However, evaluating emulators is a challenging task due to the complexity ("black box" nature) of the underlying deep models, so it is important to be able to analyze the inner functioning of these models (e.g., to analyze the predictor–predictand patterns learned) to contextualize the results. Not including this type of analysis may lead to an incomplete assessment of the emulation capabilities of both PP and MOS approaches. Some recent studies have explored the application of explainable artificial intelligence (XAI) in conventional downscaling tasks to, e.g., assess the physical consistency of the predictor patterns used by the models for inference (González-Abad et al. 2023; Baño-Medina 2020; Balmaceda-Huarte et al. 2024; Rampal et al. 2022). This new evaluation dimension allows for a better understanding of the capabilities and limitations of CNNs and may allow us to interpret how the different GCM and RCM biases may affect both emulation approaches.

One of the most promising aspects of emulators is that they could be applied to complete the scenario–GCM–RCM matrix from partial simulations, ideally from a single scenario–GCM pair. Ideally, the emulator should capture the regional climate dynamics of the RCM and be transferable. This means that it should be able to emulate other periods, emission scenarios, or even driving GCMs than those considered in the training phase. However, this remains an open issue that affects the practical use of RCM emulators for climate change applications (see, e.g., Rampal et al. 2024; Molina et al. 2023; Sun et al. 2024).

In this work, we assess the transferability of PP and MOS deep RCM emulators based on state-of-the-art CNNs. To this aim, we combine both standard and new XAI-based evaluation techniques, which allows us to measure the trustworthiness of the emulators while deepening into the understanding of the transferability of each approach to other time periods or emission scenarios (soft transferability), or driving GCMs (hard transferability). Based on the current state of knowledge, we conclude by giving some perspectives for future applications of these methods in problems where they can facilitate progress as an alternative (or complementary) to RCMs.

2. Data and methods

a. *Study area and datasets*

This study focuses on a region covering the Alps (Fig. 1), a mountain range situated in Central Europe. The distinctive features of this area enable us to assess the ability of emulators to replicate specific fine-scale atmospheric processes well reproduced by RCMs, such as the effects of topography and coastal temperature gradients.

Following previous work (Doury et al. 2023; Boé et al. 2023), we use the ALADIN63 RCM simulations over Europe at a spatial resolution of 0.11°, provided by EURO-CORDEX (Jacob et al. 2020). In particular, we use the historical experiment (1980–2005) and the RCP8.5 scenario (2006–2100), thereby encompassing a wide range of climatic conditions. These simulations are available at the Earth System Grid Federation (ESGF) for a total of four driving GCMs: NorESM1-M (Bentsen et al. 2013), CNRM-CM5 (Voldoire et al. 2013), MPI-ESM-LR (Müller et al. 2018), and HadGEM2-ES (Martin et al. 2011). Here, we consider the first three, to which we will hereafter refer by the names of their respective modeling institutions for brevity: NorESM, CNRM, and MPI. Note that HadGEM2 was not included in the study because the other three models

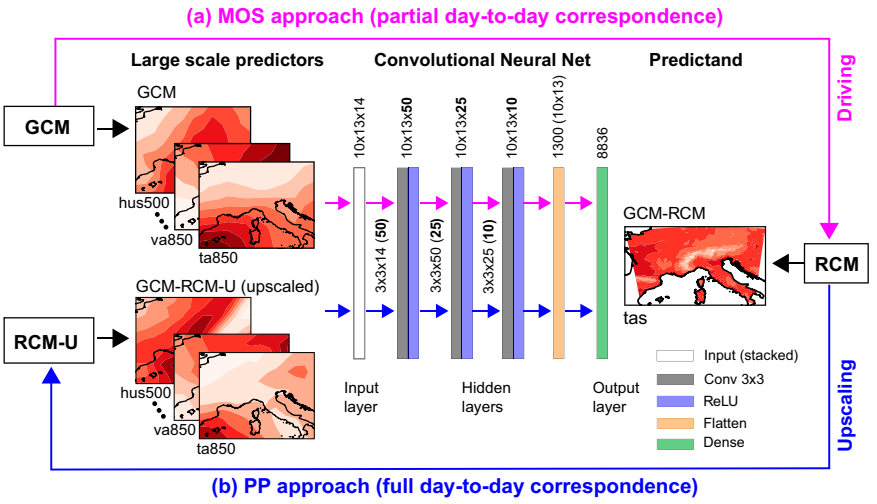

FIG. 1. Schematic representation of the DL emulator training workflows for the (a) MOS and (b) PP approaches. Arrows indicate the workflow. Details on the CNN model used are included in the figure: numbers on the top of the convolutional layers represent their size (e.g., $10 \times 13 \times 14$ in the input layer indicates the latitude, longitude, and variable dimensions, respectively), numbers between layers represent filter size, and numbers in parentheses indicate the number of filters used. The term RCM-U refers to the upscaled GCM–RCM fields.

were already downloaded at the time of the study and were sufficient to illustrate the key messages of the manuscript.

b. *Predictors and predictand*

For the predictands, we use daily fields of near-surface air temperature from ALADIN63, driven by the corresponding GCM under both historical and RCP8.5 scenarios. In this study, we only consider the land area, resulting in 8836 predictand locations/grid boxes on the 0.11° grid.

For the predictors, we selected 14 daily mean atmospheric variables which have been typically used as predictors in many downscaling applications (Brands et al. 2013; Gutiérrez et al. 2013, 2019; Baño-Medina et al. 2020; Quesada-Chacón et al. 2022): geopotential height (500 and 700 hPa), specific humidity, air temperature, and both zonal and meridional wind velocities at three different pressure levels (500, 700, and 850 hPa). For the sake of comparison, we regridded the predictor datasets using a first-order conservative remapping to a common spatial resolution of 1.5°, which is representative of CMIP5 GCMs, and is also a consensus among the GCMs utilized in this study. This leads to N samples of three-dimensional (latitude, longitude, and variable) predictor fields of dimensions $10 \times 13 \times 14$, where N is the number of days in the dataset.

The predictors were standardized at a grid-box level, using the mean and standard deviation of the training dataset.

c. *Emulation approaches*

We analyze and compare two approaches introduced in the literature for RCM emulation, termed as imperfect and perfect (Boé et al. 2023). In this study, we adhere to well-established terminology in statistical downscaling and refer to them as "MOS-emulator" and "PP-emulator," respectively, based on their connection to the MOS and PP downscaling approaches (see Maraun and Widmann 2018, for more details).

Figure 1 provides a schematic illustration of both approaches. Note that hereafter we use the terms MOS-emulator and MOS (PP-emulator and PP), indistinctly. This figure also includes a schematic diagram of the CNN model used to establish the relationship between the predictands and the predictors (see section 2e).

The MOS-emulator approach aims to learn the relationship between the target variable of the RCM and the set of predictors that are directly taken from the driving GCM (Fig. 1a, pink lines). The main shortcoming of this approach is the marginal temporal correspondence between predictors and predictand fields, since the RCM is driven by the GCM at the boundaries of the domain, but develops its own dynamics within the domain.

Differently, the PP-emulator (Fig. 1a, blue lines) aims to learn the model describing the relationship between the target variable of the RCM and the set of predictors obtained from upscaling the RCM fields to a lower resolution (RCM-U in Fig. 1). Thus, both predictors and the predictand are physically consistent and have perfect day-to-day correspondence (Doury et al. 2023).

Note that a key difference between PP and MOS emulation approaches is the potential strength of the day-to-day correspondence of predictors and predictand used to train the models. This is assessed in the results section using daily correlation as a simple and intuitive metric, i.e., how changes in the predictors over time explain (correlate with) changes in the target predictand fields (see section 3a).

d. *Experimental framework*

The experimental evaluation framework comprises two phases: training (including cross validation) and transferability. Following the methodology established by Doury et al. (2023),

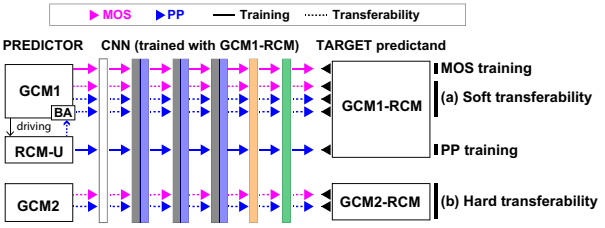

FIG. 2. Experimental framework used to assess (a) soft and (b) hard transferability (dashed lines) using the models resulting from training (solid lines). Colors indicate the approach (consistent with Fig. 1: blue for PP and pink for MOS). Note that in the case of soft PP transferability, the GCM1 predictors can be used with or without BA relative to the predictors used for training (RCM-upscaled predictors, represented as RCM-U instead of GCM1–RCM-U for simplicity).

methods are trained using both the historical period (1996–2005, historical experiment) and a far-future period (2090–99, RCP8.5 scenario), as illustrated in Fig. 1. This allows us to separate the extrapolation capability related to the stationarity assumption from our analysis (Hernanz et al. 2022; Doury et al. 2023), which is actually more relevant for standard ESD methods, as they rely on observations to project into the future (Baño-Medina et al. 2022).

The diagram in Fig. 2 represents the training of both MOS and PP methods (solid lines), using predictors from GCM1 and upscaled GCM1-driven RCM (GCM1–RCM-U), respectively. The models resulting from training are used to test transferability (dashed lines), considering both soft (changing the period and/or the scenario) and hard (changing the GCM) transferability. For convenience, soft transferability is tested in this work considering an independent midfuture period (2041–50) far away from the training periods. Soft transferability is required to fill the gaps in GCM/RCM matrices of existing climate projections, whereas hard transferability is required for emulating the RCM for new GCMs.

Note that PP methods used upscaled GCM–RCM predictors (RCM-U) for training. Then, we first test soft transferability applying the model to the GCM predictors of the same driving GCM (GCM1). However, large-scale discrepancies (biases) between GCM and upscaled RCM predictors may influence the emulated fields (Doury et al. 2023). To overcome these discrepancies, there is the possibility to bias adjust (BA) the GCM fields using as reference the upscaled RCM variables (Boé et al. 2023); these two alternatives are illustrated in Fig. 2. Bias-adjusted GCM predictors are obtained by a simple monthly adjustment of mean values relative to GCM1–RCM-U.

In hard transferability, the models trained with predictors from GCM1 (MOS) and GCM1–RCM-U (PP) are applied to the predictors from a new GCM (GCM2) and, in principle, the emulator is expected to reproduce the GCM2–RCM output; note that this is challenging due to the different biases affecting the training and emulation phases (GCM1 and GCM2, respectively). Bias adjustment is not straightforward in this case due to the different biases involved.

Note that predictors are standardized throughout all phases of the evaluation process: soft or hard transferability. This means that we consistently scale the testing predictor fields based on the mean and standard deviation of the training series.

e. Deep learning models

DL models have achieved impressive results in many data-driven applications in the last decade and had revolutionized a number of fields, including weather and climate (see, e.g., Watson-Parris 2021). DL models aim to fit a set of coefficients to a set of predictor–predictand pairs by optimizing a loss function (e.g., mean squared error) by means of the gradient descent method and the backpropagation algorithm. This ultimately means that the coefficients are progressively driven toward values of lower error in the loss surface, with a step regulated by the learning rate [see Goodfellow et al. (2016) for more details about DL]. In this work, we use CNNs, a specific type of DL models which are able to automatically infer complex spatial patterns from the input fields. In particular, we use the implementation proposed in Baño-Medina et al. (2020), known as DeepESD. This model has been successfully used for downscaling purposes in the European continent for both precipitation and temperature fields (Baño-Medina et al. 2020, 2022). Additionally, DeepESD has been analyzed with XAI techniques and proved able to learn plausible and coherent predictor–predictand links when trained with observational data (Baño-Medina 2020). Explainability is a key element of this study, used to understand the advantages and limitations of the two RCM emulator approaches.

DeepESD is a CNN composed of three convolutional layers (of 50, 25, and 10 filter/feature maps, respectively) followed by a single dense layer; the particular configuration used is shown in Fig. 1. We use "zero-type" padding, i.e., we add rows/columns of 0s to preserve the spatial dimensions after each convolutional operation. Each convolutional layer is followed by a set of rectified linear units (ReLUs) to allow the emulator to learn complex nonlinear atmospheric predictor patterns. The feature maps of the last hidden layer are flattened to build a dense connection with the output neurons, which correspond to the 8836 land grid boxes of ALADIN63 fields over the Alpine domain. The input layer is a stacked 4D (sample day^{-1}, latitude, longitude, and variable) predictor field, which feeds the hidden structure of the CNN.

On the more technical side, we use an Adam optimizer (Kingma and Ba 2014) to minimize the mean squared error between the model outputs and the RCM (groundtruth), with a batch size of 100 and a learning rate of 1E-4. The resulting DL models (i.e., one per emulator approach and GCM) contain 11.515.521 training parameters. We lean on a single NVIDIA Tesla V100 GPU with 32 GB of memory to perform both calibration and inference, with computation times of the order of a few minutes.

We perform early stopping with a patience of 30 epochs by randomly separating 10% of the training data as our validation dataset to avoid overfitting. That is, when the loss in the validation dataset does not decrease in the next 30 epochs,

then the network stops training. Note that DeepESD contains hyperparameters (e.g., number of layers or filter maps per layer) that were optimized in an extensive intercomparison study of deep learning topologies for conventional downscaling (i.e., using observational sources of records) by means of grid search (Baño-Medina et al. 2020).

f. Explainable artificial intelligence (XAI)

Neural networks are seen as black boxes due to the complex operations occurring in their hidden layers, hindering interpretability and raising distrust in the results these models produce. To overcome this aspect, several XAI techniques have been recently developed to gain understanding about the underlying patterns inferred by deep neural networks (Došilović et al. 2018). This is key to the task explored in this study, where the community aims to develop trustworthy and consistent RCM emulators to replace physics-based models. Particularly, we lean on saliency maps, which are spatial representations of the relevance of the input features to the model predictions.

The relevance of a variable is predominantly measured by computing the gradients of the output space relative to the input space. These gradients are back-propagated through the hidden layers of the network and visually displayed in the form of saliency maps. Here, we follow previous work in climate-related applications (Kondylatos et al. 2022; González-Abad et al. 2023) and compute the saliency maps using the integrated gradient (IG) algorithm (Sundararajan et al. 2017). This method integrates the gradient along the path between the input x and a baseline x'. In this study, we employ as baseline an array filled with zeros (see Mamalakis et al. 2023), representing the climatological or mean values of the standardized predictors fed into the model.

In alignment with earlier studies (Toms et al. 2021; Mamalakis et al. 2022), we postprocess the "raw" output of the XAI technique to enable a comparison across samples per day. To compute the saliency map of a particular day, we follow four steps (Toms et al. 2021; Mamalakis et al. 2022; González-Abad et al. 2023): 1) we take the absolute value since we are interested in the relevance of the features regardless of their sign. Then, 2) we compute the percentage for each saliency map individually by dividing the value of each feature by the total sum of all features. To avoid gradient shattering, 3) we filter out the lowest values using a threshold of 1.5E-3, which corresponds to 0.15% relevance. After removing the lowest gradients, the saliency maps no longer sum up to 100%. Therefore, 4) we recompute the percentage of each saliency map to ensure consistency.

Finally, we aggregate the saliency maps by averaging the values over the training period, resulting in a collection of maps with the same dimensions as the input features, which are 4D arrays in our case, each representing a predictor variable. These saliency maps can be computed for each predictand grid box. In this study, we illustrate the changing relevance patterns of the spatial predictors by focusing on four predictand grid boxes out of the 8836 covering the predictand domain. These grid boxes are located over France, the Southeastern Alps, Sardinia, and Poland, representing different behaviors and climates within the study domain.

3. Results

a. Predictor–predictand correlation

As already mentioned, a key difference between MOS and PP emulation approaches is the potential strength of the day-to-day correspondence of predictors (from the driving GCM for MOS or from the RCM for PP) and predictand (RCM near-surface temperature) used to train the models.

Figure 3 shows the daily temporal correlations between the (driving GCM or RCM) predictor fields and the predictand for the NorESM-driven ALADIN63 experiment (similar results are obtained for other driving models). In particular, we computed the correlation of surface temperature at each RCM grid box with the full predictor fields, retaining the maximum values—usually corresponding to a nearby predictor grid box—as the best estimate of the local potential link strength. We removed the seasonal cycle (monthly means) before computing the correlations, which were calculated separately for winter (DJF, columns 1–2) and summer (JJA, columns 3–4).

As expected, in general, upper-air, large-scale temperature fields show the highest correlation with near-surface temperature. Moisture (hus850) shows also widespread high correlations, while winds show lower correlations and a seasonal pattern. This is consistent with, e.g., westerlies (positive ua850) advecting relatively warm air from the ocean in winter and cool air in summer.

Apart from the overall correspondence of the different predictors with the target variable, predictors from the GCM (second and fourth columns) show systematically lower correlations than those from the upscaled RCM (first and third columns). To aid in the comparison, the maximum correlation attained and the spatial mean are shown for each panel. Moisture in summer is the only exception to the degradation of the predictor–predictand relationship in the GCM. The overall pattern with a higher correlation over northeastern Europe appears shifted to central Europe in the GCM predictors.

These results reveal that, in general, the predictors from the PP approach are more informative for the deep emulator methods used in this work than those from the MOS approach. However, the latter may have some advantages since the deep model is learning directly the relationship between the driving GCM and the target RCM variable. In this way, the different biases can be accounted for directly, in analogy to the MOS approach for statistical downscaling (Maraun and Widmann 2018).

b. Training

The different models are trained using historical (1996–2005) and far-future RCP8.5 (2090–99) simulations, considering as predictors the upscaled RCM large-scale fields (PP approach) or the GCM fields (MOS approach) and using the corresponding RCM near-surface temperature as predictand.

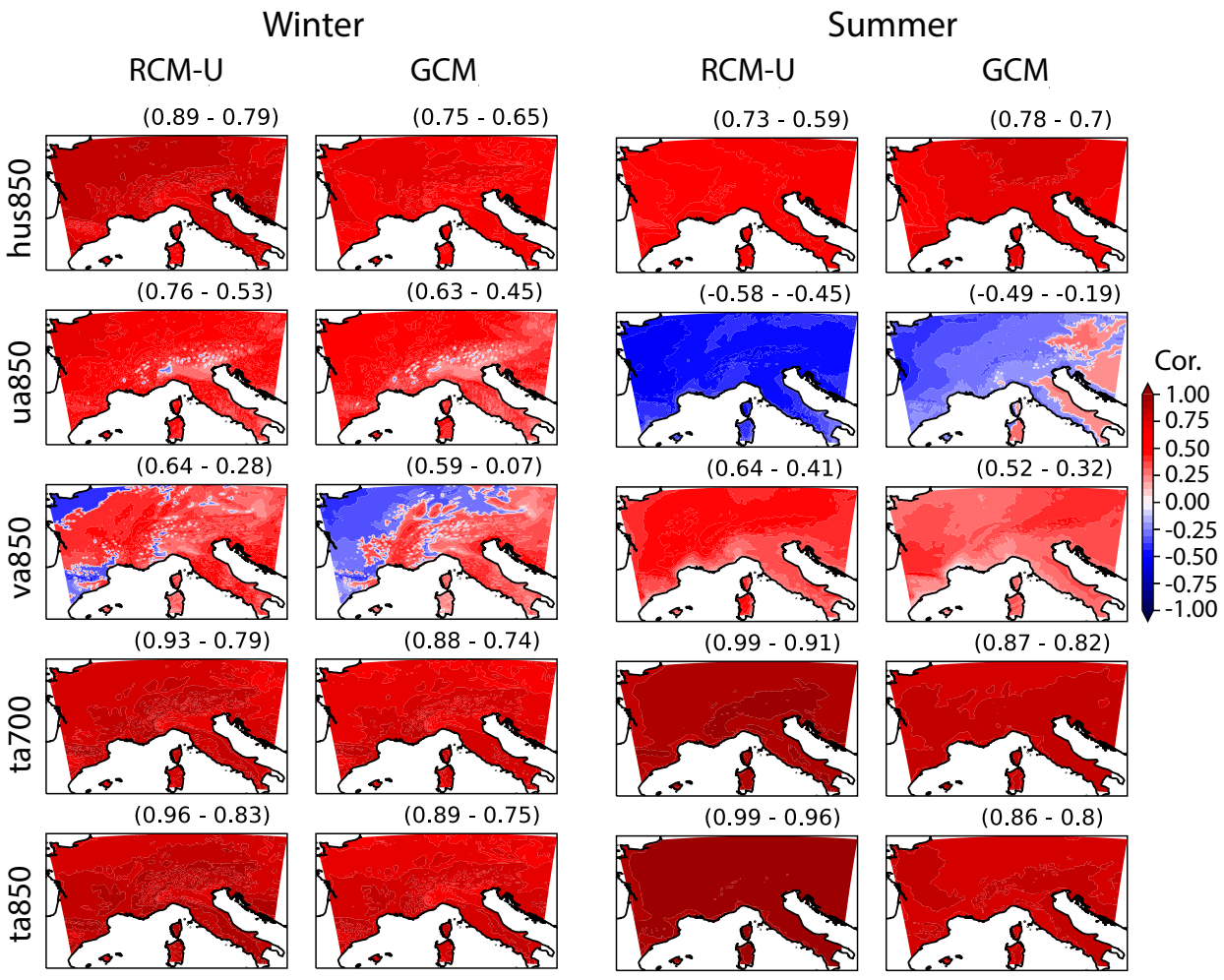

FIG. 3. Pearson temporal correlation of the daily series of the NorESM-driven ALADIN63 surface temperature (predictand) and different predictors (in rows) from ALADIN63 (RCM-upscaled, RCM-U) and the NorESM GCM for (left) winter and (right) summer. The correlation value at each grid box represents the maximum value obtained from the correlations computed between the surface air temperature (predictand) in the particular grid box and all grid boxes of the predictor fields. These values represent the best estimate of potential link strength, typically corresponding to a nearby predictor grid box. The numbers on the top of each panel indicate the spatial maximum and mean values, respectively.

Note that training is conducted using a cross-validation internal approach implemented for early stopping.

Figure 4 shows the evaluation metrics (annual and seasonal biases and RMSE, in rows) for emulated surface temperature over the training period for the ALADIN63 (RCM) simulations driven by the NorESM model (GCM). The first two columns show the training results for the MOS and PP approaches.

The PP approach attains lower RMSE values (around 1°C) than the MOS approach (2°–3°C) due to the higher correlations exhibited between predictors and predictand (see section 2a). Annual biases are generally low, but, for the MOS approach, they are the result of an average of larger seasonal biases of opposite signs in many locations. Note that the emulators are trained on an annual basis, including no specific predictor for the annual cycle.

c. *Soft transferability*

The last three columns in Fig. 4 show the results corresponding to soft transferability using predictors from the same GCM for a new middle-future period: 2041–50. For the case of the PP approach, results are included using raw (fourth column) and bias-adjusted (last column) predictors.

The results show RMSE correlation patterns remarkably similar for the PP and MOS approaches (due to the lower correlation of GCM and RCM predictors). Annual and seasonal biases are larger, particularly for the PP case, which exhibits different spatial patterns. As we show below (see Fig. 5), this occurs due to the biases between the training upscaled RCM and their counterpart test GCM fields, as already identified in previous studies Doury et al. (2023), Boé et al. (2023). Therefore, these biases could be alleviated by adjusting the GCM predictor biases (GCM-BA) relative to the upscaled RCM

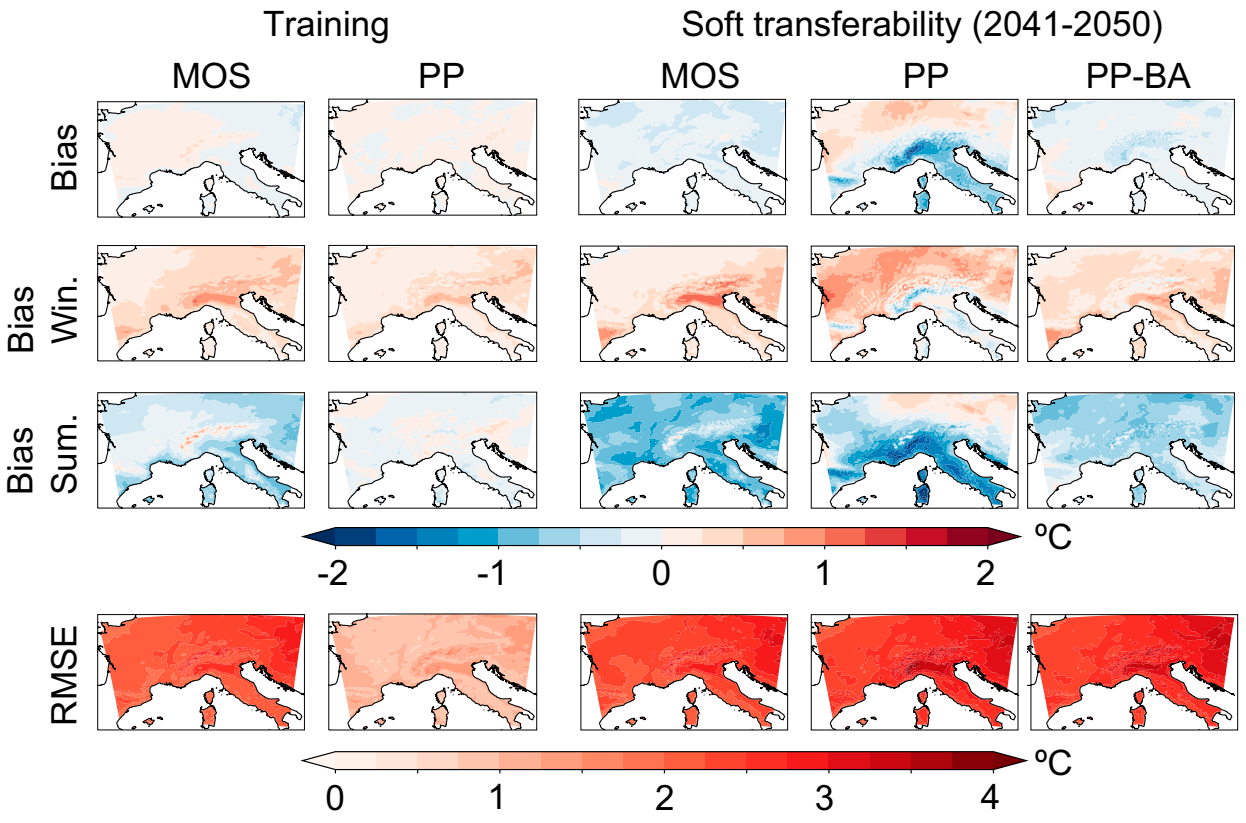

FIG. 4. Evaluation metrics for emulated surface temperature for the ALADIN63 (RCM) simulations driven by the NorESM model (GCM). The first two columns show the training results for the MOS and perfect approaches, respectively. The last three columns show the results corresponding to soft transferability using GCM-BA predictors in the last column for the PP approach. (from top to bottom) Annual, winter (DJF), and summer (JJA) biases and RMSE.

fields (RCM-U). Indeed, using GCM-BA as input (last column in Fig. 4), the emulator shows smaller biases, overall slightly smaller than those of the MOS approach (third column in Fig. 4).

Qualitatively similar results are obtained for the emulators trained on the other GCM-driven simulations. We illustrate this through Fig. 5 and the large-scale temperature (ta850), as is the most informative predictor in this study (see section 3a). Note that for other variables, such as precipitation, the interpretation of the biases would be obscured by the combined effect of several relevant predictors on the target variable. The diagram of Fig. 5 helps in better understanding the role that biases play, by representing linear transfer functions between two ordinate axes displaying the spatial averages of the predictor (ta850) and the predictand (tas), respectively. For instance, in the case of the NorESM-driven simulations, Fig. 5 shows the bias (b1) of the GCM (1) relative to the corresponding upscaled RCM predictor (2) that was used to train the EMU1 model and how this bias is reflected in the bias of the predictand (b2) when comparing the resulting model output (3) with the target RCM values (4). Results for CNRM5-driven simulations are also shown in Fig. 5 (labels 5–8 and b3–b4). Note that the two GCM predictors exhibit opposite biases when compared to the corresponding RCM upscaled predictors (b1 and b3). These biases are preserved to a large extent in the resulting predictions when comparing the emulated and actual RCM signals (b2 and b4, respectively). The maps on the sides display the corresponding spatial biases, reinforcing the idea that the large-scale biases in the predictors are inherited by the emulated surface temperature. Thus, these biases are reduced when adjusting the biases of the GCM predictors relative to the upscaled RCM fields, as indicated by the black and gray dots on the right "*y*" axis, corresponding to the results emulated from bias-adjusted GCM predictors, which are closer to the target RCM values.

These results illustrate the effect of model biases in emulators and show that MOS and PP-BA approaches seem to be suitable for emulating RCM outputs. This makes these models suitable for emulating RCM results in new periods and for new scenarios, at least when the new predictors fall into the range of variability used for learning.

d. Hard transferability: Emulating from new GCMs

Hard transferability tests whether an emulator trained with a particular GCM-driven simulation can emulate the output that the RCM would have when driven by a different GCM. In this case, the model trained using data from GCM1–RCM

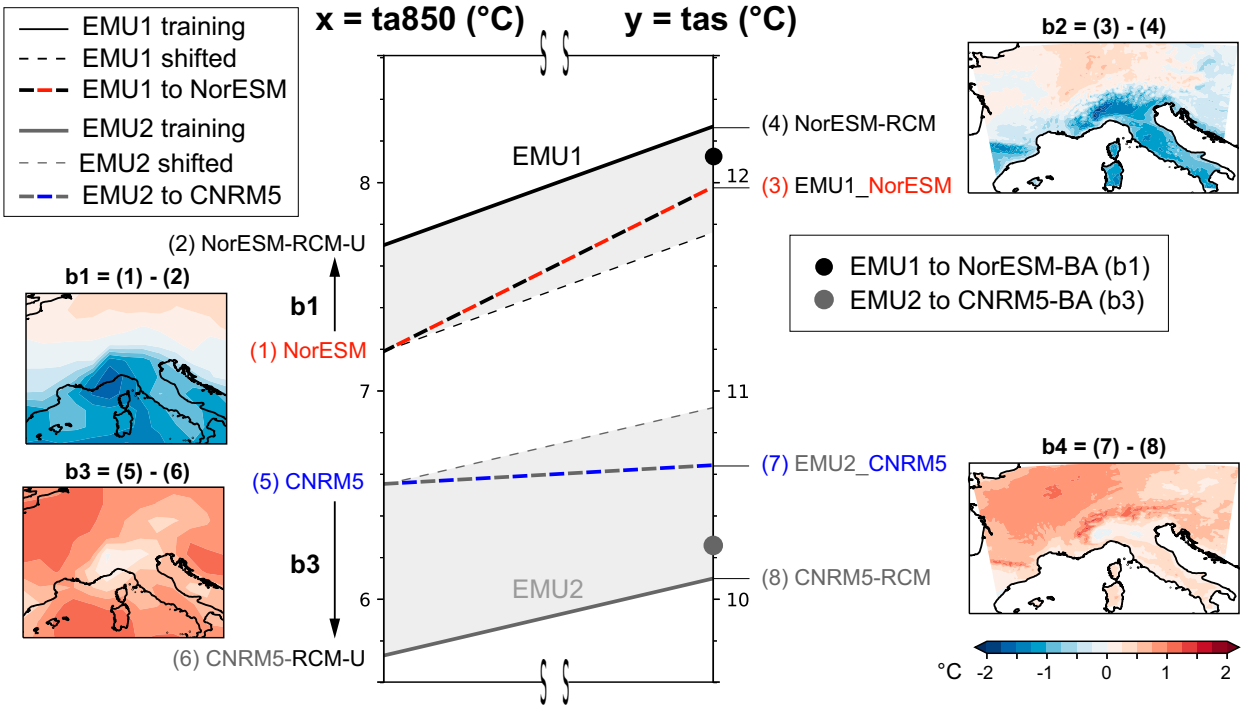

FIG. 5. Soft transferability of the perfect approach for NorESM and CNRM GCM driving simulations. The left axis shows the spatial mean values of the large-scale temperature (the "ta850" predictor) for the upscaled RCM predictors used to train the emulators (RCM-U; 2 and 6) and their driving GCMs (1 and 5). The right axis shows the spatial means of the near-surface mean temperature (the "tas" predictand) resulting from the emulated models using GCM predictors (3 and 7) and the target RCM values (4 and 8). The solid black and gray lines are linear representations of the ta850-tas relationship learned by the emulator for the NorESM (EMU1) and CNRM5 (EMU2), respectively. The thin dashed lines represent simple emulator extrapolated results for the GCM predictors (shifted EMU1 and EMU2). The different biases are indicated by b1–b4 notations and are accompanied by their corresponding spatial map representations. The results when bias-adjusted predictors are considered in the testing phase are represented by solid black and gray dots on the right axis.

is applied to a different GCM (GCM2) with the goal of reproducing the GCM2–RCM target (see Fig. 2d).

Figure 6 shows the biases resulting from hard transfer experiments (in columns) corresponding to the different combinations of different pairs of train and test GCMs from the set of available models (NorESM, CNRM, and MPI). The first two rows show annual biases for the PP and MOS methods, respectively. The other rows show the corresponding seasonal results for winter and summer. In most of the cases, the resulting biases exhibit a similar spatial pattern for the PP and MOS approaches, with smaller intensity for the former. Also, the sign of the bias reflects the influence of the GCM on the emulated fields, with opposite biases when reversing the train–test GCMs (cf. columns 1–3, 2–5, and 4–6). This could reflect that the spatial structure of emulator biases is a consequence of the different biases of the GCMs used for training and prediction.

In the case of soft transferability, the biases between the GCM and the upscaled RCM large-scale predictors were an avoidable source of error for RCM emulators. However, bias adjustment is not applicable for hard transferability, since adjusting GCM2 predictors relative to GCM1 would effectively yield a target output close to GCM1–RCM, instead of the desired target GCM2–RCM.

The key here is that the biases between the upscaled RCM and the driving GCM can have opposite signs, as is shown in Figs. 5b1 and 5b3. Thus, adjusting the biases across GCMs can be catastrophic for the emulator. Figure 7 illustrates the effect of this hard transferability on the spatial average. As in Fig. 5, emulators are depicted as linear transfer functions between two ordinate axes representing the spatial average ta850 (most informative predictor) and tas (predictand). In this case, we also depict the MOS-emulator (EMU3) trained with raw NorESM GCM predictors and NorESM-RCM tas. On the spatial average, the emulators behave close to their linear representation, indicated in the plot by a simple shift of the emulator to the new test predictors (thin dashed lines), preserving the slope. For instance, when applied to the CNRM GCM, the relatively strong slope of the MOS-emulator (10) and the slighter one of the PP-emulator (9) are preserved. But both positive slopes bring the emulator predictions far from the target (8).

Adjusting for the GCM bias would only worsen the results, since the emulator is already providing warm tas estimates

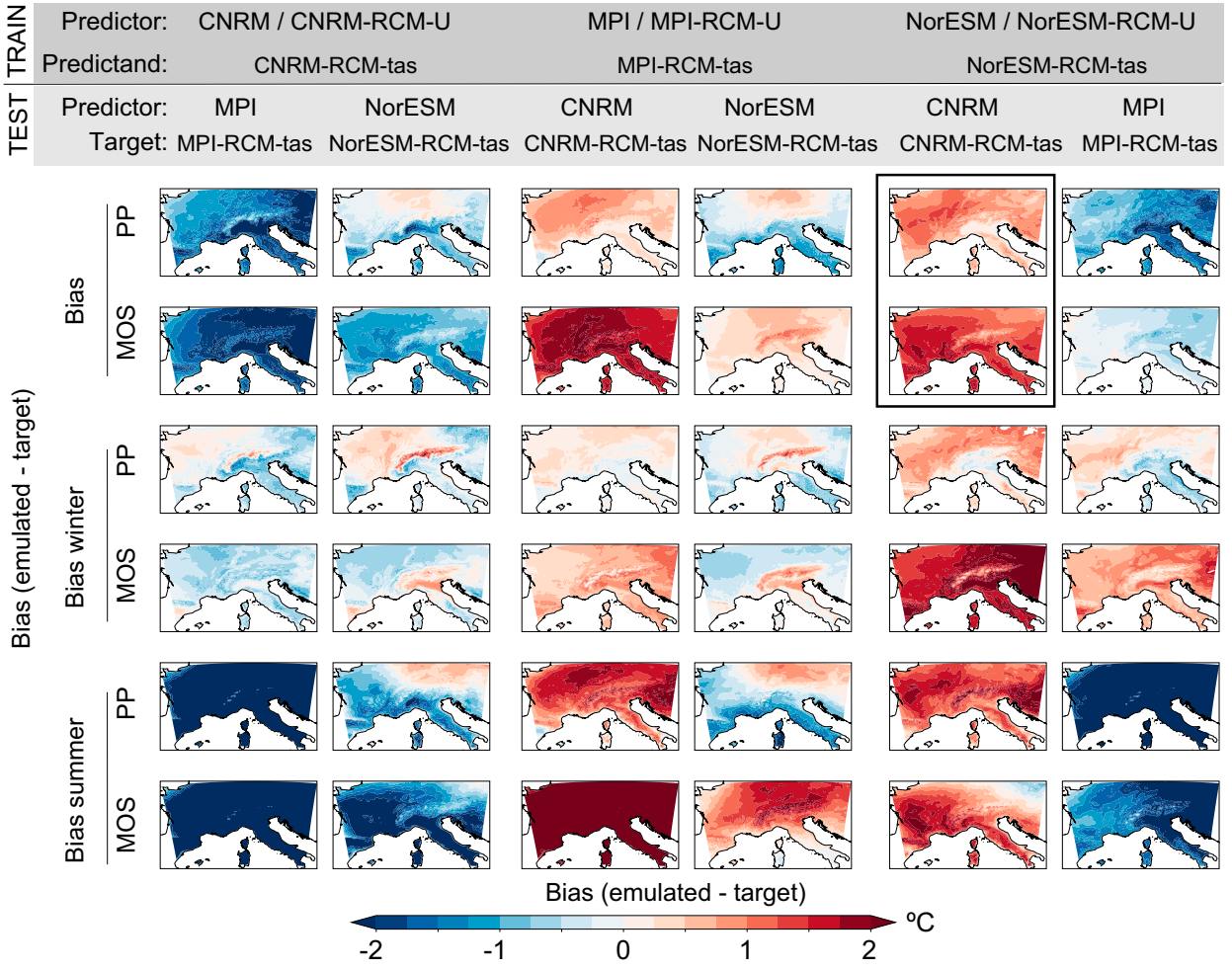

FIG. 6. Hard transferability biases considering the entire set of possible combinations (or "transfers"), that include three GCMs (CNRM, MPI, and NorESM)—which could be used either for training or testing, —and two different emulator approaches (PP and MOS). Columns 1–2 represent the bias, measured by the difference between the emulated and the target field. Differently, columns 3–4 and 5–6 are representations of the bias during the winter (DJA) and summer (JJA) seasons, respectively. We identify each of the panels by adding information about (in rows) the testing predictor dataset, (top) the predictor and predictand datasets used for training, and (in columns) the emulator approach. Note that in the top part of the figure, we use the symbol "/" to differentiate the predictor datasets used for the MOS and PP approaches. For instance, the top-left panel is the bias computed using the MPI air surface temperature as the target field, for the PP-emulator trained using CNRM-RCM-U and CNRM-RCM for the predictor and predictand datasets, respectively, and then feeding the model with MPI predictors for testing.

using colder ta850 inputs from CNRM5. If these inputs are adjusted to the warmer NorESM fields, the emulator would provide even stronger warm tas biases. For the PP-emulator, the true adjustment to apply would be b3, which accounts for the difference between the upscaled RCM and the new GCM fields (CNRM5) provided as test input. This adjustment would nudge the predicted fields in the right direction. Note, however, that this adjustment cannot be done in practice, since the aim of this hard-transferability exercise is to avoid the RCM simulation nested into a second GCM. And, if this simulation is available, it is always more worthwhile to train a new emulator (EMU2 in Fig. 7) on these new data than trying to adjust the inputs of an emulator (EMU1) trained on a different GCM.

e. *Explainability*

Figure 8 shows the aggregated (over the training period) saliency maps for two illustrative locations obtained for the PP and MOS deep learning models for two illustrative driving GCMs (NorESM and CNRM). The figure displays the saliency maps for a selection of predictors: specific humidity (hus), zonal (ua), and meridional (va) wind velocities at 850 hPa and air surface temperature (ta) at 700 and 850 hPa, in rows. The number of each panel indicates the total contribution to the output (%) of each predictor variable. The saliency maps are displayed in pairs of columns, grouping the results for the two GCMs to facilitate the analysis of the patterns learned when using different driving GCMs. Columns 1–2 (3–4) and

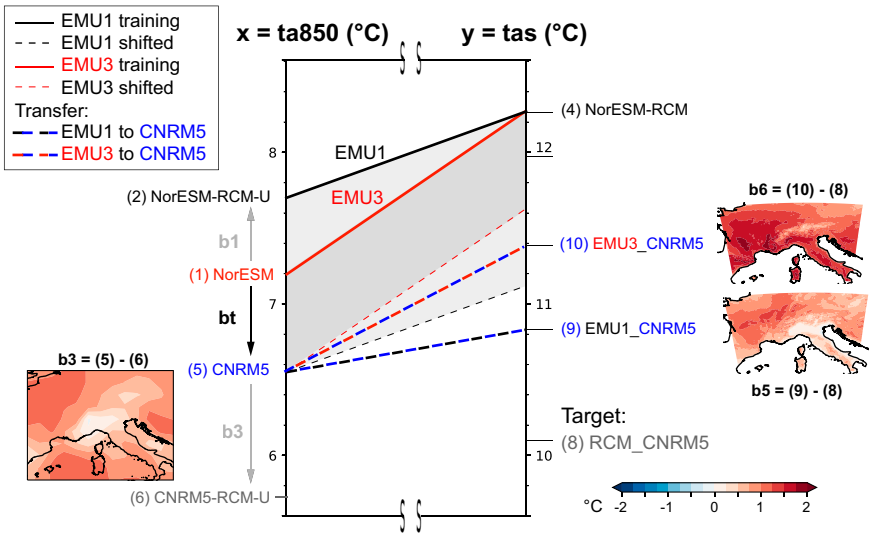

FIG. 7. An illustration of hard transferability for both PP and MOS approaches. Black and red solid lines are linear representations of the ta850–tas PP (EMU1) and MOS (EMU3) emulators trained using NorESM simulations, respectively. We then examine the hard transferability problem by feeding the NorESM PP- and MOS-emulators (EMU1 and EMU3) with data from the CNRM5 GCM (EMU1 to CNRM and EMU3 to CNRM, respectively). The left axis shows the spatial mean values of the large-scale temperature (the "ta850" predictor) for the upscaled RCM fields (RCM-U; 2 and 6) and the GCMs (1 and 5). The right axis shows the spatial means of the near-surface mean temperature (the "tas" predictand) resulting from the PP- and MOS-emulators using CNRM predictors (9 and 10, respectively) and the target RCM values (4 and 8). The thin dashed lines represent simple emulator extrapolated results for the GCM predictors (shifted EMU1 and EMU2). The different biases are indicated by b3 and b5–b6 notations and are accompanied by their corresponding spatial map representations.

5–6 (7–8) show the PP (MOS) results for two illustrative locations in the Alps and Poland, respectively.

There are remarkable differences in the relevance patterns resulting from the two approaches. PP results are very similar for the two GCMs, exhibiting a high local character, with patterns centered on the target location. Moreover, the preferred predictors and largest influence areas are physically plausible, since they are restricted mostly to the neighborhood of the target location, with a main dependence on lower-level temperatures and, to a lesser extent, on specific humidity, aligning with findings from prior studies (Huth 2004; Baño-Medina 2020). This gives high confidence in the deep learning emulation function, as it demonstrates its capability to consistently extract patterns of influence from different GCM-driven simulations of the same RCM. MOS patterns are in general more difficult to interpret, exhibiting nonlocalized or misplaced patterns that change from model to model. In this example, the results for CNRM are in better correspondence with the patterns learned with the PP approach, whereas the results for NorESM are in general more difficult to interpret, exhibiting nonlocalized or misplaced relevance patterns. This could be a consequence of the smaller day-to-day correspondence between predictors and predictand for MOS, which could result in statistical artifacts with no physical consistency during the optimization process. These results highlight the importance of introducing explainability in the evaluation of emulators, particularly for the MOS approach. The results for the MPI model (not shown) are more similar to the CNRM than the NorESM models.

These differences are further illustrated in Fig. 9 displaying the spatial correlations of the saliency maps for humidity at 850 hPa (the second most relevant predictor, following upper-air temperatures) across different locations (four illustrative locations over the area of study) and three GCMs (NorESM, CNRM, and MPI). The correlations for the different MOS/PP results are shown in the upper/lower triangles. This figure shows that the interGCM correlations for the PP approach are high for each of the locations, indicating similar spatial predictor patterns (from the upscaled GCM-driven RCM fields) being extracted by the different emulators (see the dashed boxes along the diagonal on the figure) for the same location. On the other hand, the PP results exhibit low inter-location correlations, indicating that the models learn specific spatial predictors for different locations. This is a desired behavior for the RCM emulators from a physical point of view.

Contrarily, correlations from the MOS approach are in general medium and low, both across GCMs and across locations, indicating no apparent structure in the predictor fields relevant to the different models and locations. This is particularly relevant for the NorESM model, whereas the MOS results for CNRM and MPI are closer to the PP results. Therefore, there is no guarantee that RCM emulators trained under the MOS

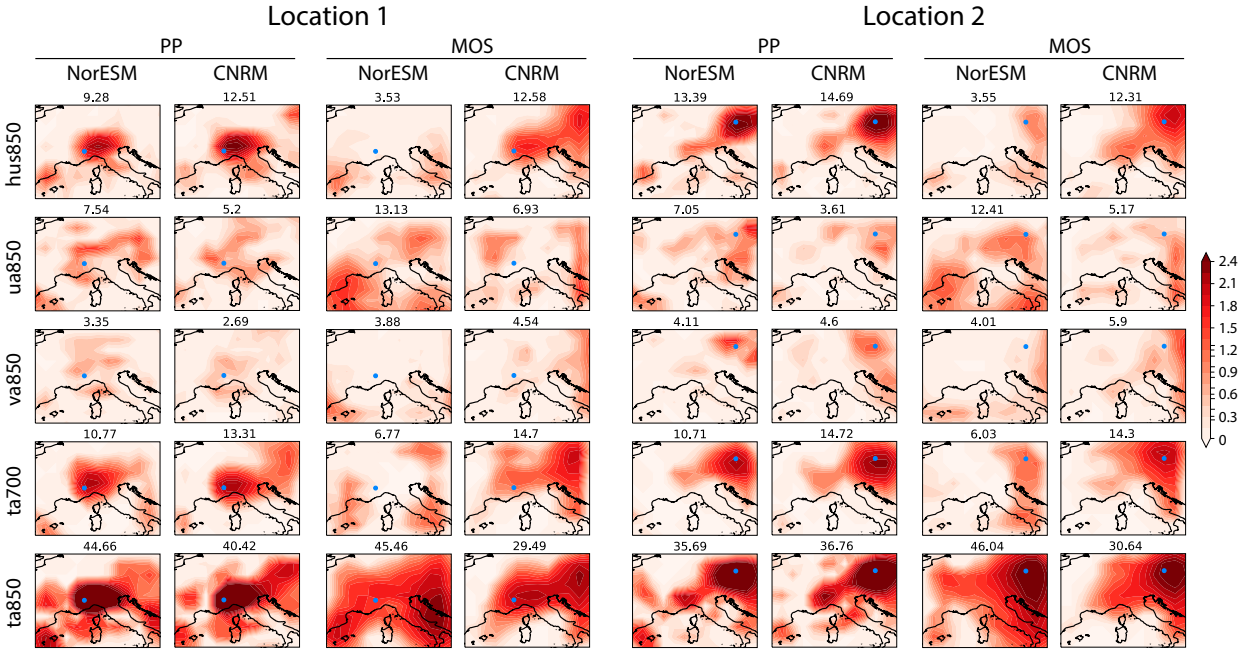

FIG. 8. Aggregated (over the training period) saliency maps for winter temperature for two illustrative locations (loc1: close to the Alps and loc2: Poland; indicated by blue dots in each panel) for the PP and MOS models learnt for the NorESM and CNRM driving GCMs. Saliency values are displayed for specific humidity (hus), zonal (ua), and meridional (va) wind velocities at 850 hPa and air surface temperature (ta) at 700 and 850 hPa, in rows. The numbers of each panel indicate the total contribution to the output (%) of each predictor variable.

approach are able to extract meaningful physical information from the predictors. This requires a case-by-case assessment involving further research building on physical principles and processes.

4. Conclusions and prospects

We examined the two most common approaches for the emulation of RCMs, "imperfect" and "perfect," building on a set of existing GCM–RCM simulations from the CORDEX initiative. First, we renamed these emulation approaches as MOS-emulator and PP-emulator, respectively, inspired by their similarity to model output statistics and perfect prognosis statistical downscaling. Second, we also coined the terms "soft transferability" and "hard transferability," which allow us to classify the different use cases for deep learning emulators. Third, besides standard validation approaches, we evaluated both MOS- and PP-emulators on the basis of physical consistency by means of an XAI technique (saliency maps), assessing their benefits and shortcomings in the different use cases. Ultimately, this type of analysis measures the trustworthiness and interpretability of the emulation function. These are key evaluation aspects, especially when the ultimate goal is to emulate a physical model.

To analyze the transferability of the emulator function to other time periods or emission scenarios of the driving GCM used for training (soft transferability), we tested both approaches over the midcentury period 2041–50. We examined the differences between the prediction and the groundtruth (biases) in both PP and MOS emulated fields. For PP, emulated fields suffer from biases that are mostly inherited by the GCM–RCM ones. We found that they can be largely reduced by means of a BA algorithm that adjusts for the differences between the GCM and RCM-U predictor fields. Differently, MOS does not inherit the GCM–RCM biases since it has been directly trained using GCM predictor fields but still shows slight biases in the emulated fields. In light of these results, we cannot clearly identify what approach is best for the emulation of an alternative temporal period. In this regard, the XAI analysis provided us with additional information that allowed us to discriminate between approaches and provide recommendations. We found that the PP approach learns predictor–predictand relationships with a strong local dependence on the low-level temperature and humidity that resembles the actual climate dynamics. This agrees with studies dealing with conventional downscaling (i.e., trained on observational datasets), suggesting similar local links for temperature (Baño-Medina 2020; González-Abad et al. 2023). On the contrary, the predictability of MOS-emulators leans on spatially extended, nonlocal patterns. These differences between the MOS and PP approaches are the result of the different correlations between the input and output training fields. Under low day-to-day correspondence, the machine learning model is more likely to find predictability sources anywhere in the spatial domain. This lack of physical meaning in the MOS predictor–predictand relationship might be driving some of the biases in the MOS emulated fields, since they are remarkably larger than the ones observed for PP during

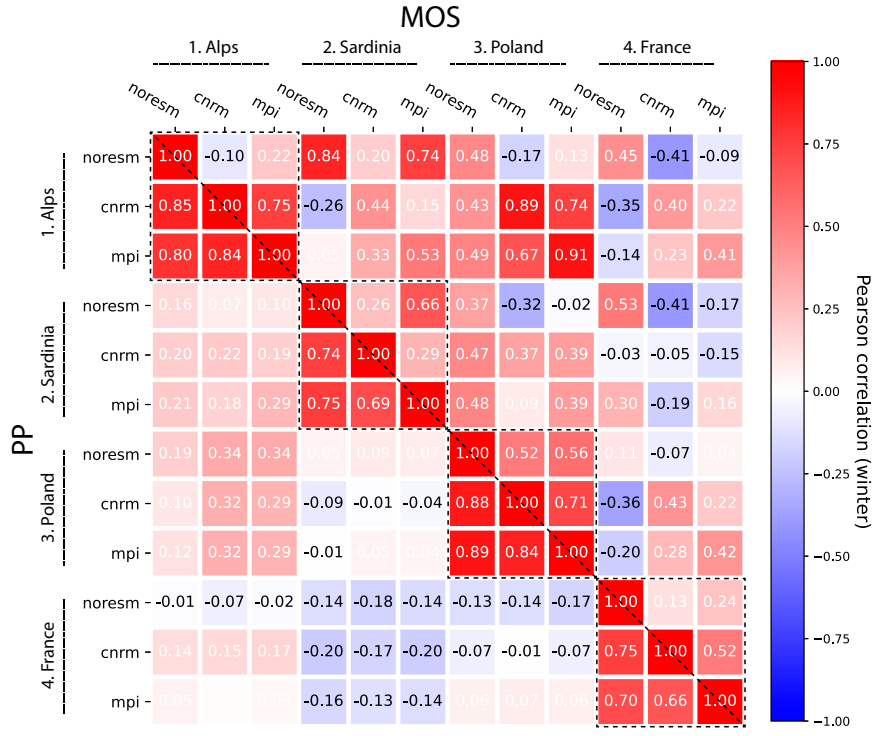

FIG. 9. Correlation of saliency maps of humidity at 850 hPa for four illustrative locations for three RCM emulators built from different driving GCMs (NorESM, CNRM, and MPI). Values in the upper/lower triangle correspond to the results for the models learned under the MOS/PP approaches (separated by the dashed diagonal). Dashed boxes along the diagonal show the interGCM results for each of the locations.

training and contrasts with the interpretability of the PP emulation function. These results suggest that the PP approach is more trustworthy and reliable than MOS for emulating alternative temporal periods of the training GCM. This contrasts with previous studies that recommended the use of MOS-emulators against PP ones on the basis of conventional evaluation metrics (Boé et al. 2023; van der Meer et al. 2023).

Another potential use case is to apply the emulator to a GCM different from the one used during training (hard transferability). In this case, we provided a thorough examination of the unique nature of the GCM–RCM biases across GCMs that lead to unreliable emulation regardless of the methodology considered, either PP or MOS. Since the RCM response to the predictors can differ greatly, even in sign, between driving GCMs, the emulator cannot foresee this response in which it was not trained. Adjusting biases across GCMs produces uncontrolled artifacts in the results and may lead to good results due to bad reasons. Nevertheless, the XAI analysis allowed us to identify the transferability of the emulation function in "ideal" case scenarios, i.e., assuming the GCM–RCM biases are independent of the GCM. We inspected the predictor–predictand relationships over a set of representative locations under this hypothetical scenario. We found that the MOS emulation function learned is very different for the different GCM–RCM pairs considered during training, meaning that it could not be applied to downscale other GCMs. However, the PP emulation function learned is very similar regardless of the training GCM, so it could be transferred to other GCMs. Given these prospects with the PP approach, further understanding of the nature of the GCM–RCM biases would be needed, to identify transferability windows of opportunity—i.e., cases where the emulator function can be used to downscale other GCMs than the ones seen during training. For instance, recent work suggests that structural differences (e.g., aerosols representation and atmospheric physics) between the RCM and driving GCM are a large driver of the dissimilarity between their large-scale fields, being almost negligible when both simulations are driven with consistent external forcing among them (Taranu et al. 2023). One such example would be the CNRM-CM5-driven simulation, which indeed shows a bit more local predictor saliency patterns under MOS training than the other GCMs. Another way forward can be either to 1) train the emulator under non-biased driving fields, which could be used later also to debias new predictors, or 2) train the emulator with a wider variety of GCM biases. The former could be accomplished by training with reanalysis data (so-called perfect boundary conditions for the RCM), but this would limit the training to the current climate, compromising the ability of the emulator to extrapolate under future climate conditions. The latter would imply mixing different driving GCMs in the training phase, to show the emulator the RCM response to different biases. This

approach might enhance the capability of the emulator to cope with different biases, but it will likely affect its accuracy for any particular GCM input as compared to a specific GCM–RCM emulator as those shown in this work and elsewhere.

Overall, we found that both approaches face problems in emulating the results for different driving GCMs, thus limiting their applicability to fill the GCM–RCM combination matrix of regional climate projections. However, there is an ongoing discussion arguing that the PP approach may potentially lead to an alternative GCM–RCM combination matrix of simulations (Doury 2022). This is based on the assumption that the RCM can be perceived as a composition of two functions: a transformation of the large scale and a downscaling function. The former arises from the fact that the RCM develops its own large-scale dynamics since it is only constrained by the GCM fields at the domain boundaries, while the latter accounts for the increase in resolution. Since the PP approach is trained to learn the downscaling function and not the large-scale transformation component of the RCM, it should not be expected to respond to the GCM fields in the same way as the RCM does. In this regard, the XAI analysis presented in this study is crucial to building trust in such an alternative GCM–RCM matrix. This is because we found that the PP emulation function learned for a given GCM–RCM pair is transferable to other GCMs in the hypothetical scenario where GCM–RCM biases are removed, which ultimately also means that the large-scale transformation of the RCM is not considered.

Emulators have also great potential for convection-permitting RCM (CPRCM) simulations (Coppola et al. 2020). These are very costly RCM simulations with grid spacings below 4 km, where the parameterization representing convection can be deactivated, reducing uncertainties, e.g., for precipitation. These simulations are typically nested into the same RCM at a coarser resolution (e.g., the 12-km resolution considered in this study). In this context, the whole emulation process is performed in the framework of the same RCM, thus avoiding the problem of the biases arising from the driving model mismatch with the emulated one. The emulator would learn the relationship between the RCM and CPRCM and provide CPRCM-emulated fields out of inputs from RCM simulations at coarse resolution.

Note that in this work we did not cover spatial transferability, i.e., training the model in one region (or in a representative selection of regions) and using it in different regions, which is also an active field of research (Bjerre et al. 2022; Ludwig et al. 2023). We also did not consider the use of spectral nudging in the RCM simulations, since these are not commonly used in future regional climate scenarios (e.g., they are discouraged in CORDEX). However, it would be of interest to conduct an intercomparison study of the MOS- and PP-emulator approaches for this type of RCM simulations in a future study, since nudging greatly reduces the large-scale transformation mentioned above, bringing the RCM-U fields closer to the GCM fields and unifying both approaches.

The emulation function is described by the neural network's coefficients, and therefore, different topologies can lead to potentially different plausible fields. In this study, we relied on DeepESD, since it is a well-tested DL topology that has proved capable of downscaling climate change scenarios over Europe in previous studies (Baño-Medina et al. 2022). In this regard, another important step toward the emulation of RCMs is to intercompare different DL topologies for both MOS and PP approaches by means of, e.g., XAI methods, as was done in González-Abad et al. (2023) for conventional statistical downscaling. This can even lead to ensembles of DL emulators that encapsulate the uncertainty on the emulation function.

Another aspect that can be further explored is the set of predictors. In this study, we followed previous studies and built on large-scale variables that are commonly employed in statistical downscaling (see, e.g., Brands et al. 2013). However, the particularities of this novel hybrid dynamical–statistical method may demand different variables than the ones traditionally used in downscaling. An example might be the dependence of RCMs climate change signal on aerosols (Taranu et al. 2023). Thus, emulators may potentially benefit from the inclusion of climate change forcings (aerosols and GHG concentrations) in the predictor set. Also, one other key aspect is the normalization of the predictor variables. In this study, we followed conventional practices in statistical downscaling, scaling each grid box independently by computing statistics (i.e., mean and standard deviation) across the time dimension (Baño-Medina et al. 2022). Recent work has explored alternative normalization methodologies, e.g., by scaling the predictors across the space dimension independently for each day, in order to preserve the daily spatial coherence (Doury et al. 2023).

Acknowledgments. This work is part of IMPETUS4-CHANGE, funded by the European Union's Horizon Europe research and innovation programme under Grant Agreement 101081555. J. M. G. and J. F. acknowledge support from MCIN/AEI/10.13039/501100011033, which funded projects ATLAS (PID2019-111481RB-I00) and CORDyS (PID2020-116595RB-I00), respectively. We would like to express our gratitude to the anonymous reviewers for their valuable feedback and insightful comments, which greatly contributed to the improvement of this manuscript. We extend our special thanks to Antoine Doury for his valuable suggestions and ideas, which clarified the methodology and were also incorporated into the conclusions of this work as potential lines of further research.

Data availability statement. Both GCM and RCM data used in this study—from the Coupled Model Intercomparison Project Phase 5 (CMIP5) and CORDEX, respectively—are openly available from the Earth System Grid Federation (ESGF, https://esgf.llnl.gov) portal. The code necessary to reproduce the results presented in this paper is available on Zenodo at the following DOI: https://doi.org/10.5281/zenodo.10966706.

REFERENCES

Balmaceda-Huarte, R., J. Baño-Medina, M. E. Olmo, and M. L. Bettolli, 2024: On the use of convolutional neural networks for downscaling daily temperatures over southern South

America in a climate change scenario. *Climate Dyn.*, **62**, 383–397, https://doi.org/10.1007/s00382-023-06912-6.

Baño-Medina, J., 2020: Understanding deep learning decisions in statistical downscaling models. *CI2020: Proc. 10th Int. Conf. on Climate Informatics*, Online, Association for Computing Machinery, 79–85, https://dl.acm.org/doi/10.1145/3429309.3429321.

——, R. Manzanas, and J. M. Gutiérrez, 2020: Configuration and intercomparison of deep learning neural models for statistical downscaling. *Geosci. Model Dev.*, **13**, 2109–2124, https://doi.org/10.5194/gmd-13-2109-2020.

——, ——, and ——, 2021: On the suitability of deep convolutional neural networks for continental-wide downscaling of climate change projections. *Climate Dyn.*, **57**, 2941–2951, https://doi.org/10.1007/s00382-021-05847-0.

——, ——, E. Cimadevilla, J. Fernández, J. González-Abad, A. S. Cofiño, and J. M. Gutiérrez, 2022: Downscaling multi-model climate projection ensembles with deep learning (DeepESD): Contribution to CORDEX EUR-44. *Geosci. Model Dev.*, **15**, 6747–6758, https://doi.org/10.5194/gmd-15-6747-2022.

Bentsen, M., and Coauthors, 2013: The Norwegian Earth System Model, NorESM1-M – Part 1: Description and basic evaluation of the physical climate. *Geosci. Model Dev.*, **6**, 687–720, https://doi.org/10.5194/gmd-6-687-2013.

Bjerre, E., M. N. Fienen, R. Schneider, J. Koch, and A. L. Hojberg, 2022: Assessing spatial transferability of a random forest metamodel for predicting drainage fraction. *J. Hydrol.*, **612**, 128177, https://doi.org/10.1016/j.jhydrol.2022.128177.

Boé, J., A. Mass, and J. Deman, 2023: A simple hybrid statistical–dynamical downscaling method for emulating regional climate models over Western Europe. Evaluation, application, and role of added value? *Climate Dyn.*, **61**, 271–294, https://doi.org/10.1007/s00382-022-06552-2.

Brands, S., S. Herrera, J. Fernández, and J. M. Gutiérrez, 2013: How well do CMIP5 Earth System Models simulate present climate conditions in Europe and Africa? *Climate Dyn.*, **41**, 803–817, https://doi.org/10.1007/s00382-013-1742-8.

Cardoso, R. M., and P. M. M. Soares, 2022: Is there added value in the EURO-CORDEX hindcast temperature simulations? Assessing the added value using climate distributions in Europe. *Int. J. Climatol.*, **42**, 4024–4039, https://doi.org/10.1002/joc.7472.

Coppola, E., and Coauthors, 2020: A first-of-its-kind multi-model convection permitting ensemble for investigating convective phenomena over Europe and the Mediterranean. *Climate Dyn.*, **55**, 3–34, https://doi.org/10.1007/s00382-018-4521-8.

Diez-Sierra, J., and Coauthors, 2022: The worldwide C3S CORDEX grand ensemble: A major contribution to assess regional climate change in the IPCC AR6 atlas. *Bull. Amer. Meteor. Soc.*, **103**, E2804–E2826, https://doi.org/10.1175/BAMS-D-22-0111.1.

Došilović, F. K., M. Brčić, and N. Hlupić, 2018: Explainable artificial intelligence: A survey. *2018 41st Int. Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, Croatia, Institute of Electrical and Electronics Engineers, 210–215, https://doi.org/10.23919/MIPRO.2018.8400040.

Doury, A., 2022: Robust estimation of regional climate change: Construction of an hybrid approach between deep neural networks and climate models. M.S. thesis, Department of Mathematics, Computer Science and Telecommunications, l'Universiťe Toulouse 1 Capitole, 197 pp., https://theses.fr/2022TOU10053.

——, S. Somot, S. Gadat, A. Ribes, and L. Corre, 2023: Regional climate model emulator based on deep learning: Concept and first evaluation of a novel hybrid downscaling approach. *Climate Dyn.*, **60**, 1751–1779, https://doi.org/10.1007/s00382-022-06343-9.

Erlandsen, H. B., K. M. Parding, R. Benestad, A. Mezghani, and M. Pontoppidan, 2020: A hybrid downscaling approach for future temperature and precipitation change. *J. Appl. Meteor. Climatol.*, **59**, 1793–1807, https://doi.org/10.1175/JAMC-D-20-0013.1.

Fernández, J., and Coauthors, 2019: Consistency of climate change projections from multiple global and regional model intercomparison projects. *Climate Dyn.*, **52**, 1139–1156, https://doi.org/10.1007/s00382-018-4181-8.

Giorgi, F., 2019: Thirty years of regional climate modeling: Where are we and where are we going next? *J. Geophys. Res. Atmos.*, **124**, 5696–5723, https://doi.org/10.1029/2018JD030094.

González-Abad, J., J. Baño-Medina, and J. M. Gutiérrez, 2023: Using explainability to inform statistical downscaling based on deep learning beyond standard validation approaches. arXiv, 2302.01771v1, https://doi.org/10.48550/arXiv.2302.01771.

Goodfellow, I., Y. Bengio, and A. Courville, 2016: *Deep Learning*. MIT Press, 800 pp., http://www.deeplearningbook.org.

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

——, and Coauthors, 2019: An intercomparison of a large ensemble of statistical downscaling methods over Europe: Results from the VALUE perfect predictor cross-validation experiment. *Int. J. Climatol.*, **39**, 3750–3785, https://doi.org/10.1002/joc.5462.

Hernanz, A., J. A. García-Valero, M. Domínguez, and E. Rodríguez-Camino, 2022: A critical view on the suitability of machine learning techniques to downscale climate change projections: Illustration for temperature with a toy experiment. *Atmos. Sci. Lett.*, **23**, e1087, https://doi.org/10.1002/asl.1087.

Hobeichi, S., N. Nishant, Y. Shao, G. Abramowitz, A. Pitman, S. Sherwood, C. Bishop, and S. Green, 2023: Using machine learning to cut the cost of dynamical downscaling. *Earth's Future*, **11**, e2022EF003291, https://doi.org/10.1029/2022EF003291.

Huth, R., 2004: Sensitivity of local daily temperature change estimates to the selection of down-scaling models and predictors. *J. Climate*, **17**, 640–652, https://doi.org/10.1175/1520-0442(2004)017<0640:SOLDTC>2.0.CO;2.

IPCC, 2022: Summary for policymakers. *Climate Change 2022: Impacts, Adaptation and Vulnerability*, H.-O. Pörtner et al., Eds., Cambridge University Press, 3–33, https://doi.org/10.1017/9781009325844.001.

Jacob, D., and Coauthors, 2020: Regional climate downscaling over Europe: Perspectives from the EURO-CORDEX community. *Reg. Environ. Change*, **20**, 51, https://doi.org/10.1007/s10113-020-01606-9.

Kendon, E. J., R. G. Jones, E. Kjellström, and J. M. Murphy, 2010: Using and designing GCM–RCM ensemble regional climate projections. *J. Climate*, **23**, 6485–6503, https://doi.org/10.1175/2010JCLI3502.1.

Kingma, D. P., and J. Ba, 2014: Adam: A method for stochastic optimization. arXiv, 1412.6980v9, https://doi.org/10.48550/arXiv.1412.6980.

Kondylatos, S., I. Prapas, M. Ronco, I. Papoutsis, G. Camps-Valls, M. Piles, M.-Á. Fernández-Torres, and N. Carvalhais, 2022:

Wildfire danger prediction and understanding with deep learning. *Geophys. Res. Lett.*, **49**, e2022GL099368, https://doi.org/10.1029/2022GL099368.

LeCun, Y., and Y. Bengio, 1995: Convolutional networks for images, speech, and time series. *The Handbook of Brain Theory and Neural Networks*, M. A. Arbib, Ed., MIT Press, 255–258.

Ludwig, M., A. Moreno-Martinez, N. Hölzel, E. Pebesma, and H. Meyer, 2023: Assessing and improving the transferability of current global spatial prediction models. *Global Ecol. Biogeogr.*, **32**, 356–368, https://doi.org/10.1111/geb.13635.

Mamalakis, A., I. Ebert-Uphoff, and E. A. Barnes, 2022: Explainable artificial intelligence in meteorology and climate science: Model fine-tuning, calibrating trust and learning new science. *Int. Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, Vienna, Austria, Springer, 315–339, https://dl.acm.org/doi/10.1007/978-3-031-04083-2_16.

——, E. A. Barnes, and I. Ebert-Uphoff, 2023: Carefully choose the baseline: Lessons learned from applying XAI attribution methods for regression tasks in geoscience. *Artif. Intell. Earth Syst.*, **2**, e220058, https://doi.org/10.1175/AIES-D-22-0058.1.

Maraun, D., and M. Widmann, 2018: *Statistical Downscaling and Bias Correction for Climate Research*. Cambridge University Press, 347 pp.

Martin, G. M., and Coauthors, 2011: The HadGEM2 family of Met Office Unified Model climate configurations. *Geosci. Model Dev.*, **4**, 723–757, https://doi.org/10.5194/gmd-4-723-2011.

Molina, M. J., and Coauthors, 2023: A review of recent and emerging machine learning applications for climate variability and weather phenomena. *Artif. Intell. Earth Syst.*, **2**, 220086, https://doi.org/10.1175/AIES-D-22-0086.1.

Molina, M. O., J. Careto, C. Gutiérrez, E. Sánchez, and P. Soares, 2022: The added value of high-resolution EURO-CORDEX simulations to describe daily wind speed over Europe. *EGU General Assembly 2022*, Vienna, Austria, European Geophysical Society, Abstract EGU22-1043, https://doi.org/10.5194/egusphere-egu22-1043.

Müller, W. A., and Coauthors, 2018: A higher-resolution version of the Max Planck Institute Earth System Model (MPI-ESM 1.2-HR). *J. Adv. Model. Earth Syst.*, **10**, 1383–1413, https://doi.org/10.1029/2017MS001217.

Phillips, N. A., 1956: The general circulation of the atmosphere: A numerical experiment. *Quart. J. Roy. Meteor. Soc.*, **82**, 123–164, https://doi.org/10.1002/qj.49708235202.

Quesada-Chacón, D., K. Barfus, and C. Bernhofer, 2022: Repeatable high-resolution statistical downscaling through deep learning. *Geosci. Model Dev.*, **15**, 7353–7370, https://doi.org/10.5194/gmd-15-7353-2022.

Rampal, N., P. B. Gibson, A. Sood, S. Stuart, N. C. Fauchereau, C. Brandolino, B. Noll, and T. Meyers, 2022: High-resolution downscaling with interpretable deep learning: Rainfall extremes over New Zealand. *Wea. Climate Extremes*, **38**, 100525, https://doi.org/10.1016/j.wace.2022.100525.

——, and Coauthors, 2024: Enhancing regional climate downscaling through advances in machine learning. *Artif. Intell. Earth Syst.*, **3**, 230066, https://doi.org/10.1175/AIES-D-23-0066.1.

Rummukainen, M., 2016: Added value in regional climate modeling. *Wiley Interdiscip. Rev.: Climate Change*, **7**, 145–159, https://doi.org/10.1002/wcc.378.

Soares, P. M. M., and R. M. Cardoso, 2018: A simple method to assess the added value using high-resolution climate distributions: Application to the EURO-CORDEX daily precipitation. *Int. J. Climatol.*, **38**, 1484–1498, https://doi.org/10.1002/joc.5261.

Sun, Y., K. Deng, K. Ren, J. Liu, C. Deng, and Y. Jin, 2024: Deep learning in statistical downscaling for deriving high spatial resolution gridded meteorological data. *Syst. Rev.*, **208**, 14–38, https://doi.org/10.1016/j.isprsjprs.2023.12.011.

Sundararajan, M., A. Taly, and Q. Yan, 2017: Axiomatic attribution for deep networks. *Proc. 34th Int. Conf. on Machine Learning*, Sydney, New South Wales, Australia, PMLR, 3319–3328, https://proceedings.mlr.press/v70/sundararajan17a.html.

Taranu, I. S., S. Somot, A. Alias, J. Boé, and C. Delire, 2023: Mechanisms behind large-scale in consistencies between regional and global climate model-based projections over Europe. *Climate Dyn.*, **60**, 3813–3838, https://doi.org/10.1007/s00382-022-06540-6.

Toms, B. A., E. A. Barnes, and J. W. Hurrell, 2021: Assessing decadal predictability in an Earth-system model using explainable neural networks. *Geophys. Res. Lett.*, **48**, e2021GL093842, https://doi.org/10.1029/2021GL093842.

van der Meer, M., S. de Roda Husman, and S. Lhermitte, 2023: Deep learning regional climate model emulators: A comparison of two downscaling training frameworks. *J. Adv. Model. Earth Syst.*, **15**, e2022MS003593, https://doi.org/10.1029/2022MS003593.

Voldoire, A., and Coauthors, 2013: The CNRM-CM5.1 global climate model: Description and basic evaluation. *Climate Dyn.*, **40**, 2091–2121, https://doi.org/10.1007/s00382-011-1259-y.

Watson-Parris, D., 2021: Machine learning for weather and climate are worlds apart. *Philos. Trans. Roy. Soc.*, **A379**, 20200098, https://doi.org/10.1098/rsta.2020.0098.