Pre-training for Deep Statistical Climate Downscaling: ~~A case study within the Spanish National Adaptation Plan (PNACC)~~Enhancing Consistency and Robustness Across Regional Datasets

Jose González-Abad[1], Maialen Iturbide[1], Alfonso Hernanz[2], and José Manuel Gutiérrez[1]

[1]Instituto de Física de Cantabria (IFCA), CSIC-Universidad de Cantabria, Santander, Spain
[2]Spanish Meteorological Agency (AEMET), Madrid, Spain

Correspondence: Jose González-Abad (gonzabad@ifca.unican.es)

Abstract.

Deep Learning (DL) has recently emerged as a promising approach for statistical climate downscaling. In this study, we investigate the use of pre-training in this context, building on the DeepESD model developed for the Spanish National Adaptation Plan (PNACC), which uses ERA5 predictors and the 5km ROCIO-IBEB national gridded predictand dataset. We evaluate the effectiveness of different fine-tuning strategies to adapt this pre-trained model to alternative ~~regional predictand datasets, specifically a point-based station dataset~~national and regional station (point-based) datasets. The objective is to develop downstream downscaling methods that maintain consistency with the original national-scale model while capturing the specific characteristics of regional and local datasets.

We analyze the benefits of fine-tuning~~in terms of faster convergence, improved generalization, and greater consistency~~, focusing on the improved consistency and robustness of the resulting models. Using eXplainable Artificial Intelligence (XAI) techniques, we examine the relationships learned by the models and compare the resulting climate change signals. Our results demonstrate that pre-training provides a robust foundation for statistical downscaling, particularly in cases with limited spatial and/or temporal data availability (e.g., local high-resolution datasets available only for short periods), thereby reducing epistemic uncertainty and improving the reliability of future climate projections. Overall, this approach represents a step toward standardizing DL-based downscaling models to ensure more coherent and consistent climate projections across national and regional scales.

1 Introduction

Global Climate Models (GCMs) simulate the spatio-temporal evolution of the climate by numerically solving the physical set of equations that govern its dynamics (Chen et al., 2021). GCMs are used to generate future projections along different forcing or greenhouse gas emission scenarios, providing possible future socio-economic pathways (Eyring et al., 2016). However, due to inherent physical and computational limitations, the resulting projections have a coarse spatial resolution, which limits their suitability for regional studies. Statistical downscaling (Maraun and Widmann, 2018) addresses this limitation by employing

statistical models to learn the relationship between coarse large-scale variables (predictors) and the local variable (predictand) of interest (Gutiérrez et al., 2019).

Recently, Deep Learning (DL) (Goodfellow et al., 2016; Prince, 2023) has emerged as a powerful tool for statistical downscaling, thanks to its ability to model non-linear relationships and effectively process spatial data. As a result, DL techniques have been applied to a wide range of statistical downscaling problems, from simple super-resolution (Vandal et al., 2017; Sha et al., 2020a, b) and bias adjustment (François et al., 2021) methods, to more sophisticated Perfect Prognosis (PP) (Baño-Medina et al., 2020, 2022) and emulation (Doury et al., 2023, 2024; Baño-Medina et al., 2023) approaches which rely on large-scale synoptic predictors—reliably simulated by GCMs—to learn empirical relationships with regional or local variables of interest.

Deep PP downscaling methods (hereafter deep downscaling) have already been used to produce regional climate change projections in various regions (Baño-Medina et al., 2021, 2022; Soares et al., 2023; Balmaceda-Huarte et al., 2024). For example, the new generation of regional climate change scenarios for the Spanish National Adaptation Plan, based on CMIP6 (Escenarios-PNACC 2024), includes a deep learning downscaling method (DeepESD) developed using ERA5 predictors and a 5 km gridded observational dataset over Spain (González-Abad and Gutiérrez, 2025). These downscaled scenarios serve as the primary source of information for developing impact assessments and adaptation studies in Spain. However, other regional scenario datasets have also been produced for specific applications or sub-regions, using alternative high-resolution grids or point-based station data with different downscaling techniques (Monjo et al., 2016; Amblar-Francés et al., 2020; Miró et al., 2021; Hernanz et al., 2022). Such methodological diversity~~can produce divergent outcomes , which~~ , arising from limitations in the techniques used to construct these datasets or from restricted temporal or spatial coverage, can lead to divergent outcomes that may confuse end users. The possibility of using a baseline downscaling model that can be adapted to new ~~regional datasets (high-resolution grids or point observations) would facilitate the~~ target datasets (e.g., higher-resolution gridded products, station/point observations, or alternative data sources) would improve the overall consistency of model outputs while facilitating the generation of downstream scenario products~~and improve their consistency~~. In addition, having a model with pre-learned relationships could be especially valuable in low-data regimes, which are common when working on specific regional scenarios where the amount of available data is often limited.

One promising approach to achieve this is to rely on pre-training (Bengio et al., 2006; Vincent et al., 2010; Erhan et al., 2010). In this paradigm, a DL model is first pre-trained on ~~an initial dataset(s)~~ one or more source datasets and then fine-tuned on one or several target datasets. ~~This allows the model to learn meaningful, general-purpose representations during the pre-training phase, which can then be effectively transferred to tasks related to those of the original dataset~~Pre-training and fine-tuning have been successfully applied in a wide range of domains, most commonly to improve performance on downstream tasks. For instance, much of the recent success of large language models can be attributed to ~~an~~ extensive pre-training ~~phase involving~~ on massive amounts of data (Radford, 2018; Kenton and Toutanova, 2019; Brown, 2020)~~. These pre-trained models are then fine-tuned for specific tasks, such as following human instructions accurately~~, followed by task-specific fine-tuning. Similarly, in computer vision, pre-training has ~~been instrumental, enabling fine-tuning of models that provide rich representations of spatial features~~ enabled models to learn rich spatial representations that transfer effectively across tasks (Dosovitskiy, 2020;

Radford et al., 2021; Caron et al., 2021; He et al., 2022). More recently, in weather and climate science, large deep models have been trained ~~with large combinations of datasets to provide foundational models that capture a wide range of physics-based phenomena inherent in these systems~~ on diverse datasets to capture a broad range of physical phenomena (Nguyen et al., 2023; Lessig et al., 2023; Bodnar et al., 2024; Schmude et al., 2024).

In this study, we investigate ~~, for the first time, the use of~~ pre-training for deep statistical downscaling with a different objective. Rather than focusing primarily on maximizing predictive skill for a given dataset, we use pre-training ~~in the context of deep statistical downscaling. Specifically, we~~ to encourage the learning of general-purpose representations that can serve as a stable common baseline across heterogeneous datasets, thereby improving cross-dataset consistency and robustness to dataset shifts in the resulting model outputs. During pre-training, the model learns broad, transferable structures that can later be adapted through fine-tuning to accommodate dataset-specific characteristics without substantially altering the underlying relationships.

We build on the DeepESD convolutional model originally developed for the Spanish National Adaptation Plan (PNACC), which downscales coarse-resolution GCM predictors to a high-resolution (5 km) national gridded observational dataset (González-Abad and Gutiérrez, 2025). Using this pre-trained model as a baseline, we explore the effectiveness of various fine-tuning strategies to adapt it for alternative national or higher-resolution regional predictand datasets. ~~The aim~~ Rather than seeking to maximize predictive skill for each individual dataset, the objective is to develop cost-effective downscaling ~~methods that maintain consistency~~ approaches that preserve coherence with the national-scale results while ~~capturing~~ accommodating the specific characteristics of local datasets. This is particularly relevant for the generation of national and regional climate change scenarios based on multiple observational products, as illustrated here for the regional scenarios produced within the PNACC framework. In such applications, maintaining coherent climate-change signals across products is essential for downstream impact assessments and decision-making.

We focus on daily minimum and maximum ~~temperatures~~ temperature and precipitation. To test the fine-tuning ~~performance~~strategies, we employ ~~an~~ two alternative point-based observational ~~dataset comprising over~~ datasets: one national and one regional. The national dataset comprises more than 3,400 stations for temperature and 5,800 stations for precipitation across peninsular Spain. ~~This station-based dataset serves as a~~ In contrast, the regional dataset covers the region of Catalonia and includes 114 temperature and 110 precipitation stations over a shorter temporal period, thereby reflecting the types of data limitations that the proposed methodology is designed to address. These station-based datasets constitute a compelling test case due to ~~its distinct resolution and nature compared~~ their distinct spatial resolution and observational nature relative to the original gridded data. To assess the consistency and interpretability of the fine-tuned models, we analyze the learned relationships using eXplainable Artificial Intelligence (XAI) techniques (Adadi and Berrada, 2018; Arrieta et al., 2020; Minh et al., 2022; González-Abad et al., 2023). ~~Additionally, we evaluate the benefits of pre-training in terms of faster convergence and improved generalization during the fine-tuning process.~~

The paper is structured as follows. Section 2 introduces the data, the DL model, and the XAI techniques used in this work. In Section 3, we present the pre-training and fine-tuning strategies in detail. Section 4 provides the results of all the experiments conducted in this study. Finally, in Sections 5 and 6, we discuss these results and conclude with the main findings of this work.

2 Experimental framework

In this section, we first describe the region of study, the datasets, and the preprocessing procedures. We then provide a detailed overview of DeepESD, the deep downscaling architecture that serves as the foundation for this work. Finally, we introduce the techniques used to assess the deep downscaling models, focusing on the relationships they learn.

2.1 Region of Study

We focus on peninsular Spain (36°N–44°N, 9.5°W–3.5°E), which represents a challenging benchmark for statistical downscaling due to its diverse climatology and complex orography. This region, located within the Mediterranean basin, is significantly affected by climate change, experiencing increasing temperatures and changes in precipitation patterns (Hoerling et al., 2012; Cos et al., 2022).

In this region, multiple observational datasets are available, including several gridded datasets such as ROCIO-IBEB ($5\,km$ resolution Peral García et al., 2017), Iberia01 ($10\,km$ resolution Herrera et al., 2019) and E-OBS ($10\,km$ resolution Cornes et al., 2018), as well as global higher-resolution grids ~~at 1 km (Karger et al., 2023) and~~ over specific sub-domains (Basque Government, 2020; Taboada et al., 2024). This provides opportunities to extend this work using multiple datasets of different natures for pre-training and/or fine-tuning.

2.1.1 Predictor and Predictands

As predictors, we select a set of large-scale atmospheric variables commonly used in previous climate downscaling studies (Gutiérrez et al., 2013; Baño-Medina et al., 2021; Soares et al., 2023), specifically air temperature, specific humidity, and meridional and zonal wind velocity at 850, 700 and 500hPa and mean sea level pressure. These predictors are obtained from the ERA5 reanalysis dataset (Hersbach et al., 2020) and regridded from their original $0.25°$ resolution to $1.5°$ using conservative interpolation, to match the coarser scales typical of GCM outputs.

To ensure that large-scale phenomena influencing the downscaled variables are fully captured, we extend the spatial domain to 23.5°N–68.5°N and 39°W–22.5°E. Finally, to avoid biases from differing variable scales, each predictor grid point is standardized to a zero mean and unit variance before being fed into the model.

For this study, we used ~~two~~ three types of observational data from the Spanish Meteorological Agency (AEMET) and the European Climate Assessment & Dataset (ECA&D) as predictands: the ROCIO-IBEB gridded dataset (Peral García et al., 2017), which provides daily precipitation and temperature at 5 km resolution~~, and~~ ; station observations from the STATIONS-IBEB network (Spanish Meteorological Agency, 2021), comprising over 5,800 precipitation and 3,400 temperature ground stations; and the STATIONS-CAT dataset, which denotes a specific subset of the ECA&D blended daily series (Klein Tank et al., 2002; Klok and Klein Tank, 2009), consisting of 110 precipitation and 114 temperature stations across Catalonia. ROCIO-IBEB served as the predictand for developing the deep downscaling model used to generate high-resolution gridded projections under the Spanish National Adaptation Plan (PNACC). In this study, we pre-train the downscaling model on ROCIO-IBEB and then fine-tune it on both STATIONS-IBEB ~~observations~~and STATIONS-CAT observations. Notably,

STATIONS-CAT provides an independent validation of our fine-tuning approach, as these stations were not used in the construction of ROCIO-IBEB, unlike STATIONS-IBEB. We also train ~~a separate model~~ separate models from scratch using STATIONS-IBEB ~~as the predictand~~ and STATIONS-CAT as predictands to benchmark the benefits of the ~~pre-training~~ fine-tuning approach. We refer to ~~this model~~ these models as *fully-trained* to indicate that ~~it is~~ they are trained entirely on ~~STATIONS-IBEB~~their respective station datasets, without leveraging any pre-trained model on ROCIO-IBEB.

Figure 1 presents both the climatologies of mean and extreme-related indices for each of the three predictands (minimum temperature, maximum temperature and precipitation) for ~~both~~ all datasets. Specifically, we show the mean values of the three variables as well as the annual minimum of daily minimum temperatures (TNn), the annual maximum of daily maximum temperatures (TXx), and the annual maximum daily precipitation (RX1day). Regional differences ~~in extremes —~~driven by orography and coastal versus inland conditions ~~—~~are apparent in ~~both~~ all datasets. Although the ROCIO-IBEB gridded product is derived from STATIONS-IBEB observations, the station-based dataset shows ~~higher~~ some differences in extreme values, especially for precipitation. The STATIONS-CAT dataset, despite being independent of the other two, shows similar trends and values, more aligned with those of STATIONS-IBEB. It is also worth noting that for STATIONS-IBEB the number of available stations for precipitation is higher, as reflected in the denser precipitation maps~~for the~~. For STATIONS-CAT, the number of stations is considerably lower than STATIONS-IBEB ~~dataset~~but similar between variables.

2.1.2 Historical and Future Projections

To evaluate the performance of the DL model to downscale future projections from GCMs, we follow previous studies (González-Abad and Gutiérrez, 2025) and use the EC-Earth3-Veg climate model (Döscher et al., 2021), which is among the GCMs recommended by EURO-CORDEX for downscaling CMIP6 over the European domain (Sobolowski et al., 2023), also used in Escenarios-PNACC (Correa et al., 2023). We use predictor data from the historical (1980-2014) and a future scenario (SSP3-7.0, 2071-2100) representing high emission forcing conditions. Following the assumptions of the PP approach and previous studies (Baño-Medina et al., 2021, 2022; Addison et al., 2024), we apply a simple mean-variance bias adjustment to the GCM predictors to ensure that their distribution more closely matches that of their ERA5 reanalysis counterparts. For further details on this transformation, we refer the reader to Baño-Medina et al. (2021). Prior to feeding the bias-adjusted GCM outputs into the deep-learning model, we standardize them using the ERA5 grid-box means and variances.

2.2 Deep Learning Model

We select the DeepESD architecture (Baño-Medina et al., 2022) as the basis for the standard DL model. This choice is motivated by several factors. First, while not representing the current state-of-the-art in ML-based downscaling, this model is among the most widely ~~used in the downscaling literature~~adopted for generating climate projections from GCMs, having been applied to various regions, including continental Europe (Baño-Medina et al., 2020, 2021), southern South America (Balmaceda-Huarte et al., 2024), Egypt (Kheir et al., 2023), New Zealand (Rampal et al., 2022), Germany (Quesada-Chacón et al., 2022), and, more recently, Iberia (Soares et al., 2023; González-Abad and Gutiérrez, 2025). Second, this model has been explicitly assessed for its plausibility in future climate scenarios (Baño-Medina et al., 2021; González-Abad and Gutiérrez, 2025), an aspect that is

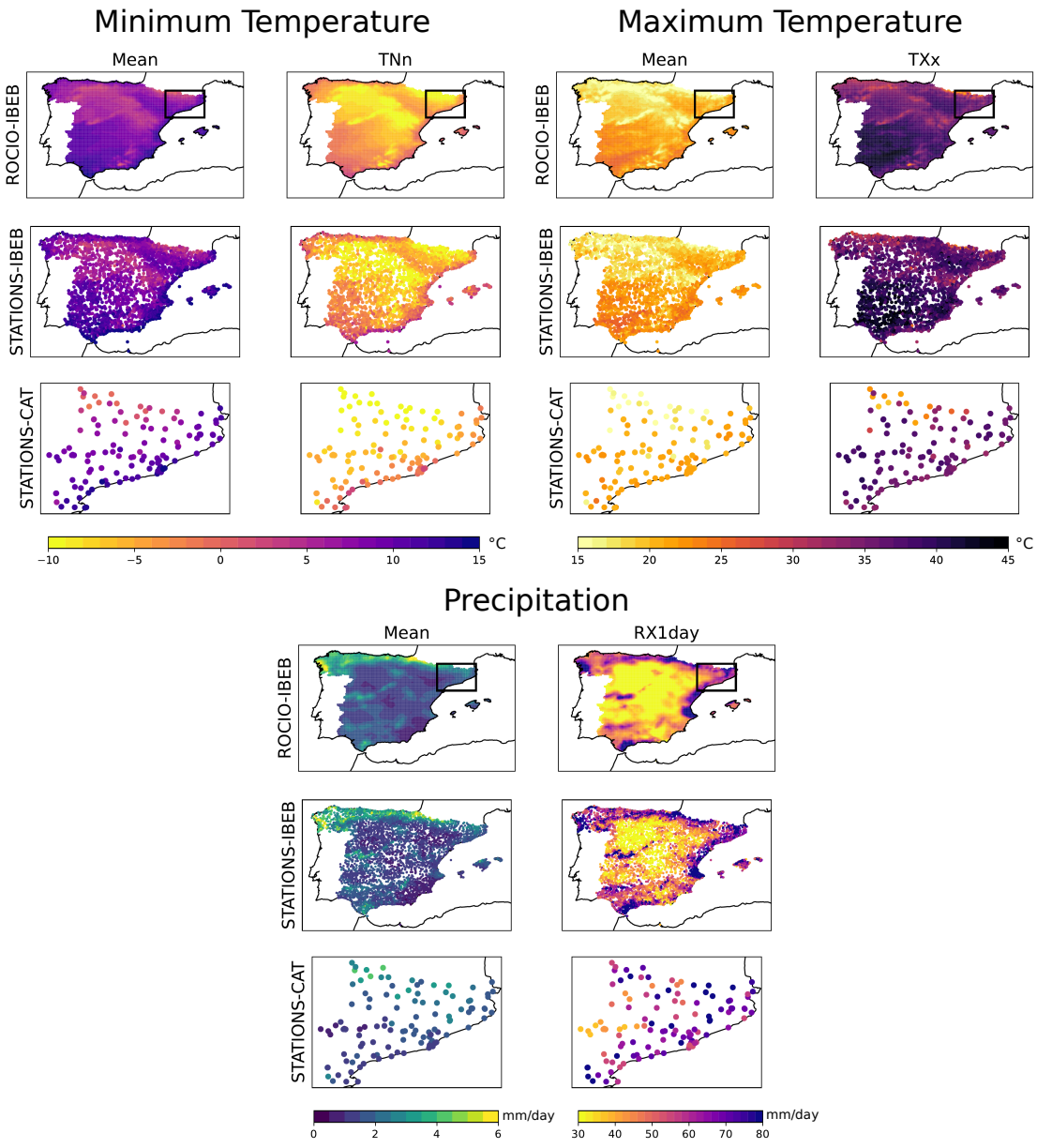

Figure 1. Climatologies ~~for the period 1980–2010~~ of minimum and maximum temperatures and accumulated precipitation. For each of the three variables, we show both the mean climatology and an extreme-related statistic: the annual minimum of daily minimum temperatures (TNn), the annual maximum of daily maximum temperatures (TXx), and the annual maximum daily precipitation (RX1day). Each variable~~'~~'s climatology is computed for the ROCIO-IBEB~~and~~ , STATIONS-IBEB and STATIONS-CAT datasets (arranged in rows within each subpanel). For the ROCIO-IBEB and STATIONS-IBEB datasets, climatologies are computed over the period 1980-2010, while for STATIONS-CAT they are computed over 2009-2018. The spatial extent of the STATIONS-CAT dataset is indicated in the ROCIO-IBEB panels.

often overlooked in downscaling studies (Rampal et al., 2024). Third, its promising results are achieved without having to rely on a complex or highly specialized architecture, making it particularly suitable for the approach explored in this study.

The DeepESD architecture employed in this work consists of three successive convolutional layers with 50, 25, and 1 kernels, respectively, each followed by a Rectified Linear Unit (ReLU) activation function (Glorot and Bengio, 2010) (see Figure 2 for a schematic overview). In the original design (Baño-Medina et al., 2020), the last convolutional layer for temperature was formed by 10 kernels instead of one. However, using 10 kernels substantially increases the network's complexity, raising the parameter count, for instance in the case of ROCIO-IBEB, from approximately 28 million (with a single final kernel) to 284 million. This increase arises from the final dense layer, which fully connects the output of the last convolutional kernel to every grid point to be downscaled. Such an overparameterized DL model may be prone to overfitting (Bishop and Nasrabadi, 2006; Hastie et al., 2001), potentially learning spurious relationships that fail to extrapolate in future scenarios, as discussed in González-Abad et al. (2023). Consequently, we choose to employ the DeepESD architecture with a single kernel in the final convolutional layer for all three variables. The output of this layer is then flattened and passed to a final dense layer, which has as many neurons as the number of points to be downscaled, 21885 for ROCIO-IBEB~~and~~ , 3460 (5803) for temperature (precipitation) for STATIONS-IBEB and 114 (110) for temperature (precipitation) for STATIONS-CAT. The differences in ~~the latter case~~ STATIONS-IBEB originate from the greater number of stations available for precipitation (see Section 2.1).

All DL models for the three variables are trained using the same procedure. We employ the Adam optimizer (Kingma, 2014) with a learning rate of 10^{-4} and a batch size of 64 to minimize the loss function. For temperature, we use the Mean Squared Error (MSE), a widely adopted loss function in temperature downscaling. In contrast, selecting an appropriate loss function for precipitation is less straightforward due to its non-continuous and exponentially distributed nature. A recent study (González-Abad and Gutiérrez, 2025) found that the asymmetric (ASYM) loss function proposed in Doury et al. (2024) is well-suited for precipitation downscaling. Consequently, we use the ASYM loss function for precipitation.

To account for variability in training performance, each model is trained ten times per variable using different random initializations of its weights. To properly evaluate the DL models, we split the dataset into training (1980–2010 for STATIONS-IBEB and 2009-2018 for STATIONS-CAT) and test sets (2011–2020 for STATIONS-IBEB and 2019-2021 for STATIONS-CAT). Additionally, during training, we set aside 10% of the training data as a validation set. This validation set is used to implement an early stopping strategy with a patience of 60 epochs, ultimately selecting the model that performs best on the validation set during this period.

2.3 Explainable Artificial Intelligence

A key factor underlying the success of DL models is their internal structure, which involves the composition of multiple piecewise/non-linear functions and a large number of parameters. However, this complexity also makes these models difficult to interpret, often branding them as *black-boxes*, as the relationships they learn are not readily apparent. This issue is particularly relevant in statistical downscaling, where the field has transitioned from simple, interpretable models to complex, deep neural networks. To address this challenge, eXplainable Artificial Intelligence (XAI) techniques have emerged, offering insights into the inner workings of DL models and shedding light on the relationships they capture (Adadi and Berrada, 2018; Arrieta

et al., 2020; Minh et al., 2022). Within the statistical downscaling domain, the application and benefits of XAI have only recently begun to be explored (Baño-Medina, 2021; Rampal et al., 2022; González-Abad et al., 2023; Baño-Medina et al., 2023; Balmaceda-Huarte et al., 2024).

In this study, we explore the relationships learned by the different DL models trained in the different regimes by applying the XAI-based diagnostics introduced in González-Abad et al. (2023)~~. These diagnostics are~~ , specifically designed for the context of PP downscaling. ~~In particular, we use~~ These diagnostics are based on saliency maps, which quantify the sensitivity of the model outputs to variations in the input predictors by measuring the gradient of the predictand with respect to the predictor. In this work, saliency maps are exploited in two complementary ways. First, they are analyzed directly by aggregating them over time, providing an average spatial representation of the saliency distribution relative to a single predictand grid point. Second, we compute the Aggregated Saliency Map (ASM), a diagnostic that ~~quantifies~~ summarizes the overall spatial influence of each predictor variable ~~'s grid points, thus indicating which points or variables the model focuses on most.~~ by aggregating saliency values across both time and all predictand grid points. This enables identification of the predictors that exert the strongest influence on the model outputs.

In this work, we compute the required saliency maps by directly calculating the gradients of the predictand space with respect to the predictor space. We then apply the same preprocessing steps to these saliency maps as in González-Abad et al. (2023) before computing the ~~ASM diagnostic~~final diagnostics.

3 Pre-training and Fine-tuning

Figure 2 presents a schematic view of the pre-training (left), partial fine-tuning (center), and full fine-tuning (right) regimes. For each regime, the corresponding DeepESD architecture is shown, with layers represented by small boxes indicating the number of neurons and kernel dimensions for the convolutional layers. The DeepESD architecture, across all training regimes, is divided into two distinct components: the feature extractor (depicted by the gray box in Figure 2) and the calibrator. The feature extractor, composed of convolutional layers, is responsible for learning high-level data representations. The calibrator, consisting of a dense layer, transforms these high-level features into localized predictions at each grid point of the predictand, enabling the model to perform point-by-point downscaling. Importantly, the architecture remains consistent across all three downscaled fields due to the design choice regarding the number of kernels in the final convolutional layer, as discussed in Section 2.2.

First, the DeepESD model is trained using ERA5 as the predictor and ROCIO-IBEB as the predictand, resulting in the *pre-trained* model shown in Figure 2. In this model, weights are randomly initialized, as indicated by the purple dots, and trained, as indicated by the blue arrows looping back to each layer. This corresponds to the standard training procedure for DL models. The pre-trained model then serves as the foundation for two additional variants: the *partial fine-tuned* and the *full fine-tuned* models. Both of these variants are trained using STATIONS-IBEB or STATIONS-CAT as the predictand, but the weights of their feature extractors are initialized using the weights learned by the pre-trained model, as shown by the purple arrows connecting the feature extractor layers across models. In the partially fine-tuned model, the feature extractor layers are

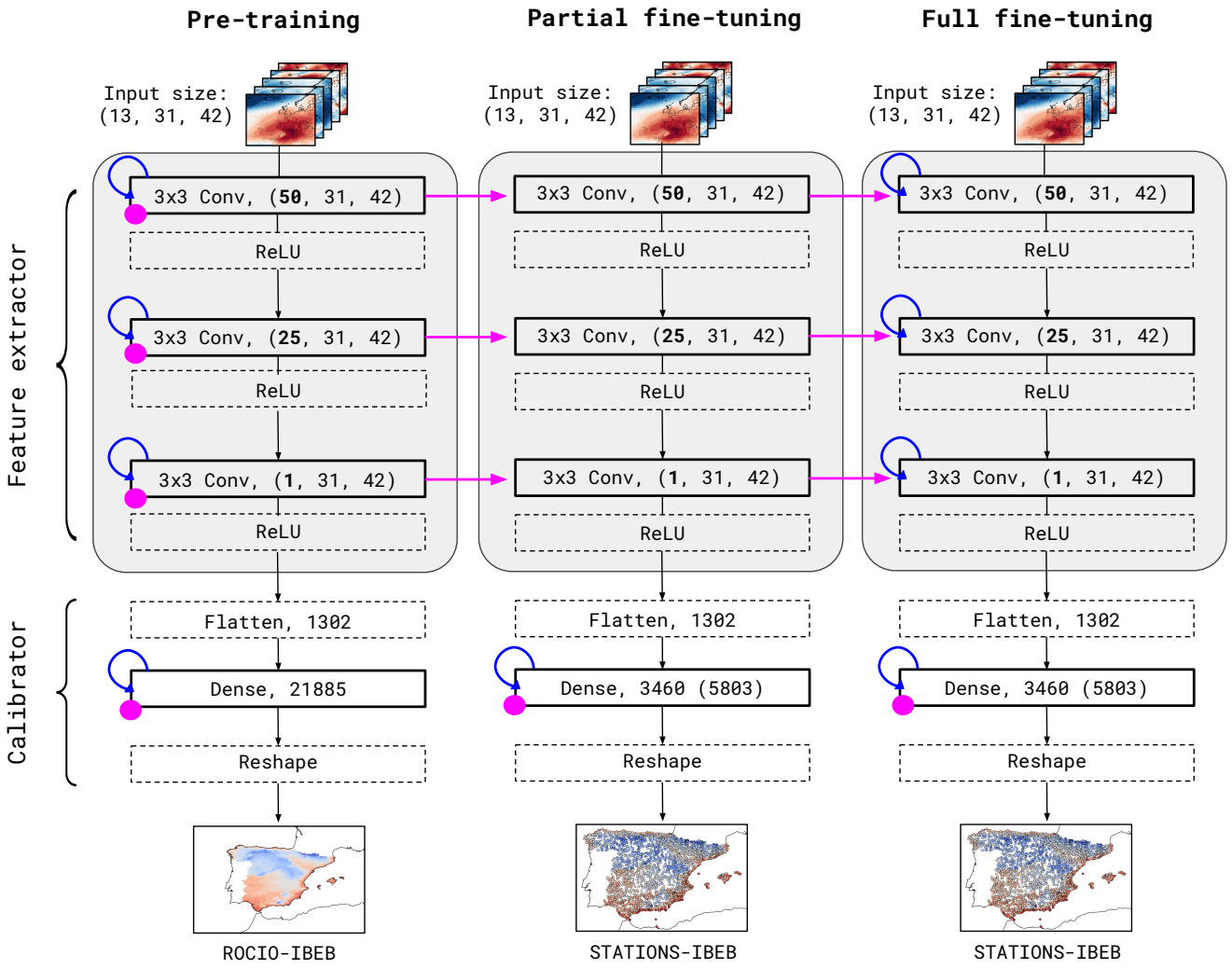

Figure 2. Schematic representation of the three training regimes: pre-training (left), partial fine-tuning (center), and ~~full-tuning~~ full fine-tuning (right). Each model is composed of convolutional and dense layers, grouped into two main blocks: the feature extractor (gray box) and the calibrator. Initialization is represented in purple, with points indicating random weights and arrows indicating initialization from pre-trained models. Blue looping arrows represent training. Layers with learnable parameters are shown with solid borders, while non-trainable layers are indicated with dashed borders. The schematic illustrates the configuration for the ROCIO-IBEB dataset. The same procedure applies to STATIONS-CAT, differing only in the number of stations represented in the final dense layer.

frozen (i.e., not updated during training), and only the calibrator is trained, which is reflected by the absence of blue arrows for the feature extractor and its presence for the calibrator. In contrast, the fully fine-tuned model also allows the feature extractor layers to fine-tune the weights inherited from the original trained model during training.

The decision to transfer only the feature extractor originates from the role that convolutional layers play in learning high-level data representations that are often transferable across related tasks (LeCun et al., 1995; Krizhevsky et al., 2012; Agrawal et al., 2014). This is particularly relevant for the DeepESD model, where the convolutional layers capture high-level synoptic patterns, while the final dense layer provides spatial specialization by fitting a linear regression over these representations for each grid point forming the predictand (Baño-Medina, 2021; González-Abad et al., 2023).

Although the feature extractor contains a relatively small fraction of the total model parameters, parameter count alone does not reflect functional importance (Goodfellow et al., 2016). The calibrator's large parameter count arises from its scaling with the number of output locations, but functionally it performs location-dependent linear mappings from the learned representations. In contrast, the feature extractor learns the nonlinear representations that determine what information is available to the final layer. In addition, prior work shows that transferability depends strongly on layer role/depth, with earlier representation layers often being more transferable than later task-specific layers (Yosinski et al., 2014).

Transferring only the weights of the convolutional layers (i.e., the feature extractor) aligns with standard practices in other domains, such as computer vision and natural language processing, where fine-tuning typically involves appending a task-specific layer to a trained backbone (or to the encoder in encoder-decoder architectures) (Devlin, 2018; Chen et al., 2020). The comparison between freezing these transferred weights and fine-tuning them reflects two prevalent strategies in the pre-training literature, where some studies keep the transferred layers fixed, while others allow them to adapt during training.

4 Results

In this section, we examine the feasibility of a pre-training strategy by comparing the performance and key characteristics of models originally trained on ROCIO-IBEB and fine-tuned on STATIONS-IBEB with those of a model fully trained from scratch on STATIONS-IBEB using randomly initialized weights. Throughout the manuscript, we refer to this latter approach as *full-training*.

4.1 ~~Performance of Deep Downscaling Models~~

~~Figure ?? shows the evolution of the loss function during training for the DeepESD model for each downscaled variable under the three training regimes for the STATIONS-IBEB dataset. As mentioned in Section 2.2, we use MSE as the loss for minimum and maximum temperatures, and ASYM for precipitation. For each regime, we show ten different training curves per combination, corresponding to the independent runs with different random initializations.~~

~~Evolution of the loss function during training for the DeepESD models across the three predictand variables (minimum temperature, maximum temperature, and precipitation) and the three training regimes (full-training, partial fine-tuning, and full fine-tuning) on the STATIONS-IBEB dataset. Each line represents one of ten training runs performed per model, using different random initializations to illustrate variability in training performance.~~

~~Overall, the fully-trained model converges more slowly, whereas the two fine-tuned models converge significantly faster. Notably, for maximum temperature downscaling, the fine-tuned models take about half the number of epochs~~ Additionally, we

present a case study using the STATIONS-CAT dataset (a regional dataset independent of ROCIO-IBEB) to demonstrate how fine-tuning enhances the robustness of climate change signals and produces more aligned and physically consistent learned relationships compared to training from scratch. ~~For all three variables, and especially for minimum temperature and precipitation, the fully fine-tuned model achieves slightly lower training errors than the partially fine-tuned model. Regarding variability across the ten training runs, the training dynamics for minimum and maximum temperature are highly consistent across all realizations. In contrast, precipitation shows greater variability.~~

4.1 Performance of Deep Downscaling Models

In Figure 3, we present the evaluation results on the test set for the four DeepESD models: the model pre-trained on ROCIO-IBEB (pre-training), the model exclusively trained on STATIONS-IBEB (full-training), and the two models fine-tuned on STATIONS-IBEB using as foundation the pre-trained model (partial and full fine-tuning). Results are shown for minimum (top) and maximum (bottom) temperature. For both variables, we report the Root Mean Square Error (RMSE) and the bias of the mean. Additionally, we include the bias in the annual minimum of daily minima (TNn) for minimum temperature, and the bias in the annual maximum of daily maxima (TXx) for maximum temperature. Note that the violin plots correspond to a randomly selected training run. To illustrate variability across model initializations, we include black dashed lines showing the minimum and maximum values of the spatial medians across the ten training replicas.

The pre-trained model ~~shows performance~~ results on ROCIO-IBEB, consistent with previous studies (Soares et al., 2023; González-Abad and Gutiérrez, 2025)~~. The~~ , are shown for reference to illustrate the characteristics of this gridded dataset. However, these results are not directly comparable to the STATIONS-IBEB models, as they involve different predictand datasets with distinct spatial characteristics. Among the three models trained on STATIONS-IBEB (~~fully-training~~full-training, partial fine-tuning and full fine-tuning)~~exhibit similar performance~~ , performance is similar across metrics, though ~~with overall lower accuracy than~~ overall accuracy is lower than for the ROCIO-IBEB model due to the more challenging nature of station-based datasets. Notably, the partially fine-tuned model displays slightly lower variability in the spatial median, particularly for the bias in extremes (TNn and TXx), suggesting improved robustness across training runs.

Figure 4 presents the equivalent analysis for precipitation. We show the Root Mean Square Error (RMSE) computed over wet days (> 1 mm/day), the bias in the frequency of dry days, the Simple Daily Intensity Index (SDII), and the bias in the annual maximum daily precipitation (Rx1day). As ~~with~~ noted for temperature, the pre-trained model ~~performance aligns~~ results on ROCIO-IBEB serve as a reference and align with previous studies~~. The~~ , but are not directly comparable to the STATIONS-IBEB models. Among the three STATIONS-IBEB models~~show comparable results~~ , results are comparable overall, though for ~~some metrics~~the bias of the SDII, the partially fine-tuned model performs slightly worse than the fully-trained and fully fine-tuned versions.

4.2 Explainability: Saliency of the Different Predictors

To analyze the relationships learned by the deep downscaling models, we use the XAI-based ASM diagnostic based on aggregated saliency (see Section 2.3 for details). Figure 5 displays this diagnostic computed over the test set for all predictor variables

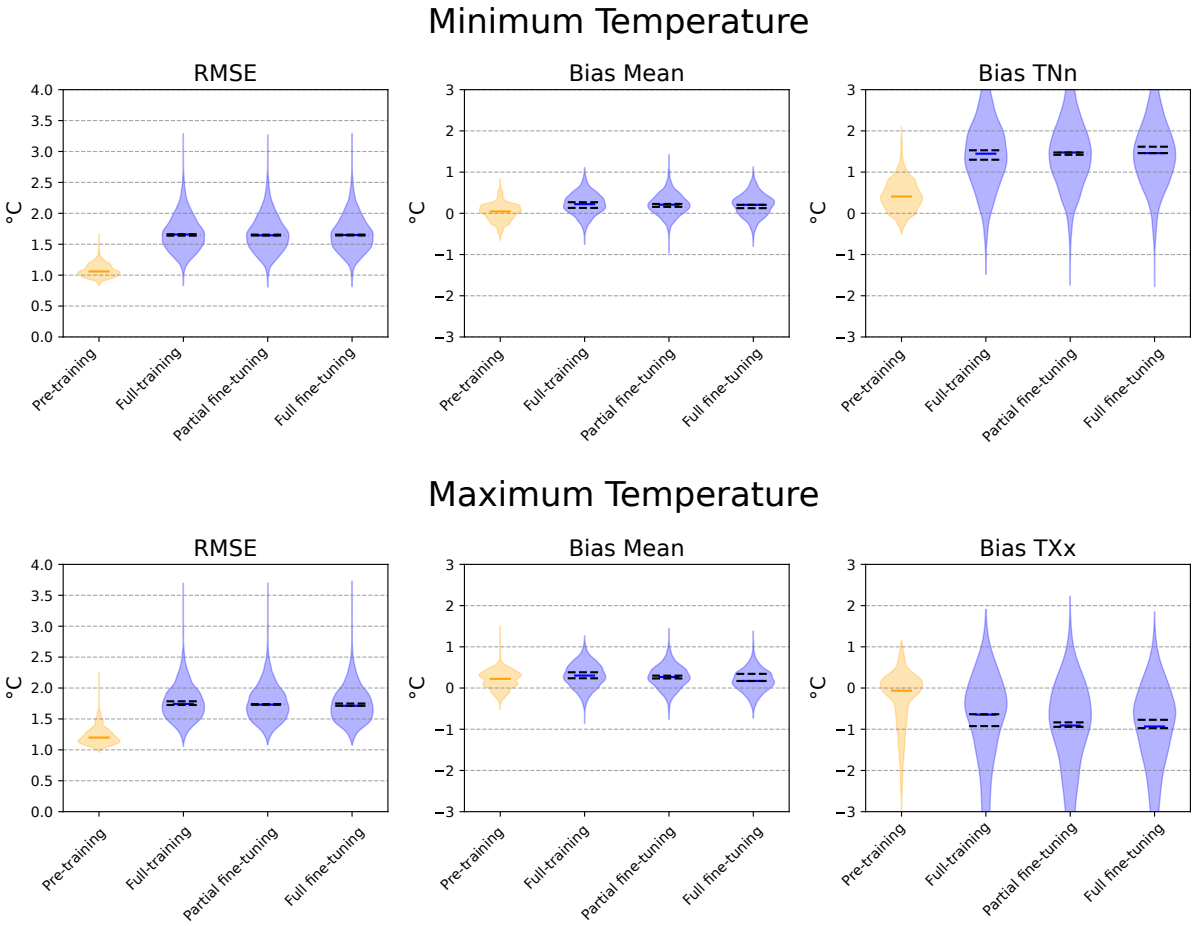

Figure 3. Evaluation results on the test set (2011–2020) for the DeepESD model trained under the different regimes~~(~~: pre-training on ROCIO-IBEB, and full-training, partial fine-tuning and full fine-tuning ~~)~~on STATIONS-IBEB. Results are shown for minimum temperature (top row) and maximum temperature (bottom row). For both variables, we report the Root Mean Square Error (RMSE) and the bias of the mean (Bias Mean). Additionally, for minimum temperature, we include the bias in the annual minimum of daily minima (TNn), and for maximum temperature, the bias in the annual maximum of daily maxima (TXx). Violin plots show the distribution across grid points for a given training run (randomly chosen), with the spatial median marked in blue. Black dashed lines indicate the range of spatial medians across the ten independent training runs.

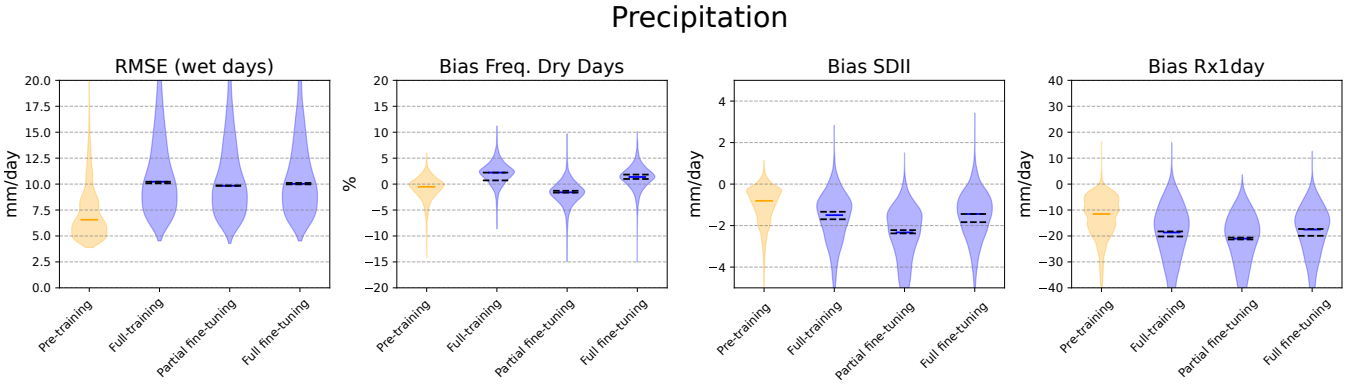

Figure 4. Same evaluation as in Figure 3 but for precipitation. Metrics include the Root Mean Square Error (RMSE) computed over wet days ($>$ 1 mm/day), the bias in the frequency of dry days (Bias Freq. Dry Days), the bias in the Simple Daily Intensity Index (Bias SDII), and the bias in the annual maximum daily precipitation (Bias Rx1day).

for the three downscaled variables (shown in separate subplots) and for the pre-trained ~~,~~on ROCIO-IBEB and fully-trained, partially fine-tuned and fully fine-tuned models on STATIONS-IBEB. In previous work (González-Abad et al., 2023), ASM values are depicted for every grid point in each predictor variable, thereby illustrating spatial patterns of relevance. However, to simplify our analysis, we aggregate the ASM spatially, showing a single value per predictor variable that represents its overall importance for the deep downscaling model. In Figure 5, these aggregated saliency values are displayed as histograms, with the color of each bar (gray, red, green, or blue) corresponding to the different models.

For minimum temperature, the ASM for both the pre-trained and the three STATIONS-IBEB models is distributed across multiple variables, with air temperature and specific humidity at 850 hPa, along with mean sea level pressure, emerging as the most relevant. However, the model trained on the ROCIO-IBEB dataset (pre-training) and the STATIONS-IBEB trained model (full-training) differ in how they attribute importance to certain variables (for example, the zonal wind component at 850 hPa). This discrepancy diminishes when the weights of the feature extractor are transferred (fine-tuning), resulting in a closer alignment with the pre-trained model. Interestingly, the full fine-tuning diverges more noticeably in its ASM, showing greater differences in variable relevance than even the fully-trained model for some predictors (e.g., mean sea level pressure). For maximum temperature, the pattern is notably different. In both the pre-trained model and the three STATIONS-IBEB models, nearly all relevance is assigned to mean sea level pressure, with only a minor contribution from air temperature at 850 hPa. In contrast, precipitation shows distributed relevance across all predictor variables, reflecting the complexity of its underlying processes. Similar to minimum temperature, certain discrepancies arise between the pre-trained and the fully-trained model for variables such as the zonal wind component and the mean sea level pressure. These discrepancies are again resolved under either one of the two fine-tuned models, where the final ASM values converge to those of the pre-trained model.

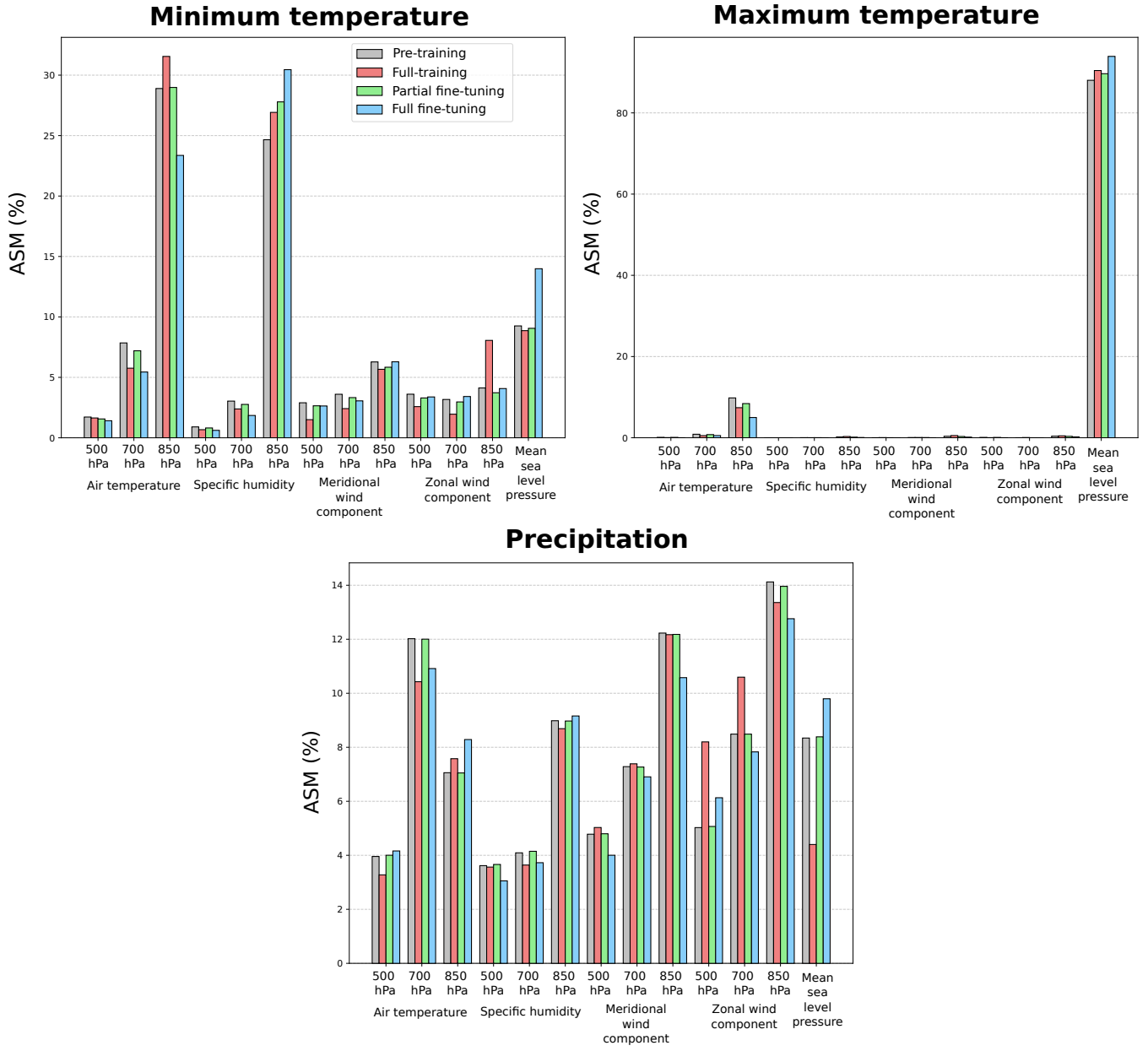

Figure 5. Aggregated Saliency Map (ASM) computed over the test set for the three downscaled variables (shown in separate subplots). The model pre-trained on the ROCIO-IBEB dataset is depicted in gray, while the three models for STATIONS-IBEB (full-training, partial fine-tuning and full fine-tuning) are shown in red, green, and blue, respectively. The ASM is spatially aggregated for each predictor variable, resulting in a single importance value per variable, as represented by the bars.

4.3 Downscaled Climate Projections

To assess the extrapolation capabilities of the models trained under the different regimes, we compute the climate change signal for the three downscaled variables from the EC-Earth3-Veg model under the SSP3-7.0 scenario (see Section 2.1.2 for details). For temperature, this signal is obtained by subtracting the downscaled future projections (2071–2100) from those of the historical period (1980–2014), while for precipitation, it is calculated by dividing the future projections by the historical ones. Figure 6 shows the resulting climate change signals for the TNn, TXx, and Rx1Day indices, which correspond to minimum temperature, maximum temperature, and precipitation, respectively. The rows represent each index, while the columns display the signals from the original GCM as well as from the pre-trained model and the three training regimes for the STATIONS-IBEB dataset (full-training, partial fine-tuning and full fine-tuning). For the temperature-related indices (TNn and TXx), all three training regimes produce climate change signals that are broadly consistent with those of the pre-trained model. However, for TNn, these signals diverge from the climate model's output, despite showing comparable magnitudes of change. In contrast, for Rx1Day, both the pre-trained model and the three training regimes capture the climate model's signal while adding regional detail, such as reduced extreme precipitation in northern Spain. Despite this overall agreement, the fully-trained model slightly underestimates the change in the Duero River basin relative to both the pre-trained and the GCM, while the fully fine-tuned model tends to overestimate it. Only the partially fine-tuned model closely reproduces the magnitude and spatial pattern of change seen in both the GCM and the pre-trained model.

To assess the variability of the resulting climate change projections, we train the DeepESD model ten times under each of the three different regimes, changing only the random initialization of the parameters. In the fully-trained model, all parameters are randomly reinitialized for each replica, whereas in the two fine-tuning regimes, only the parameters of the final dense layer are reinitialized. Figure 7 shows the standard deviation of the climate change signals for the TNn, TXx, and Rx1Day indices (rows) across model replicas for each training regime (columns). As expected, the fully-trained model exhibits the highest variability as no parameters are fixed, increasing the sources of variation across replicas. In contrast, although the fully fine-tuned model updates all parameters, it begins each training with the same initialized feature extractor, reducing variability across replicas. This effect is even more pronounced in the partially fine-tuned model, where the feature extractor is kept fixed and only the dense layer is trained. For temperature indices, both fine-tuning regimes show similar low variability. However, for precipitation, the fully fine-tuned model shows greater variability than the partially pre-trained one, particularly in southeastern Spain.

4.4 Sensitivity to ~~data record length~~Data Record Length

~~Finally, to evaluate the benefits of relying on~~ To evaluate whether the relationships learned by the pre-trained ~~models in low-data regimes~~model are beneficial when data availability is limited, we train the aforementioned DL models on different versions of the STATIONS-IBEB dataset, each featuring a distinct proportion of missing data points. These versions are created by randomly removing observations across both space and time. During training, the DL models update their parameters only taking into account the non-missing data points of the predictand, which enables them to learn even when portions of the

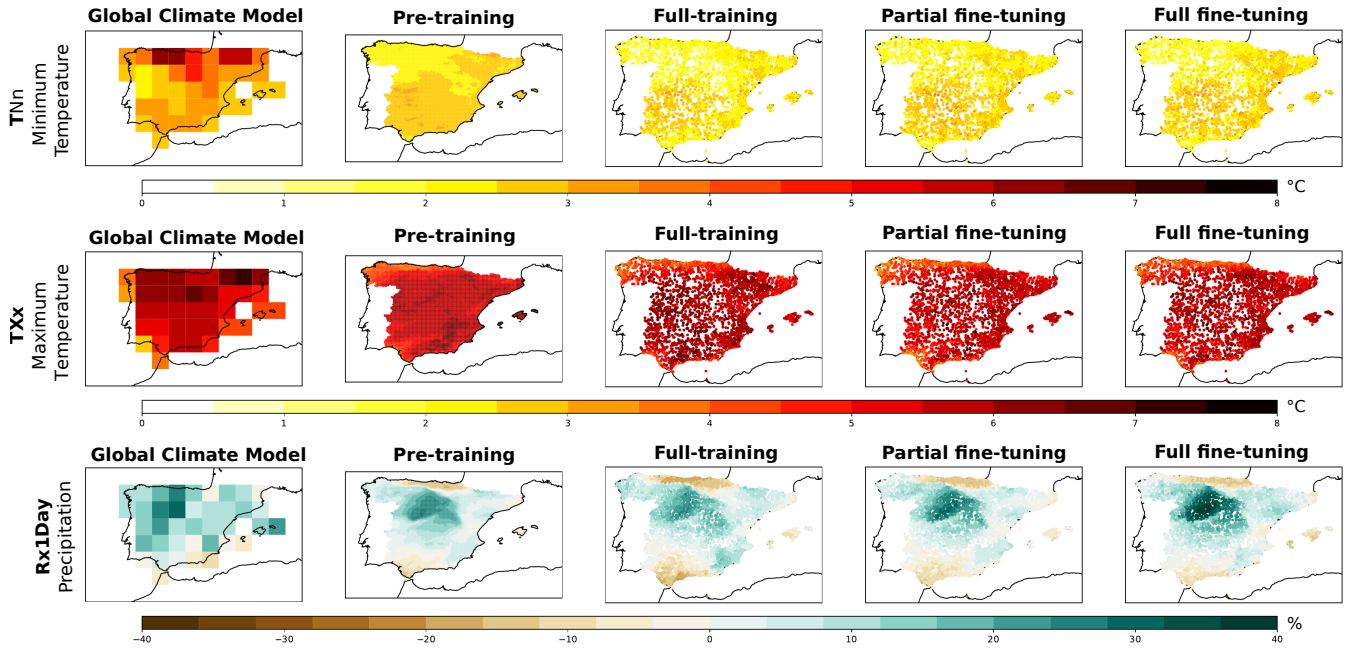

Figure 6. Climate change signals corresponding to the EC-Earth3-Veg model for the annual minimum of daily minimum temperatures (TNn), the annual maximum of daily maximum temperatures (TXx), and the annual maximum daily precipitation (RX1day). Each index is shown in a separate row, displaying the climate model's signal as well as the downscaled signals from the pre-trained model and the three different training regimes for the STATIONS-IBEB dataset (fully-training, partial fine-tuning and full fine-tuning), in columns. Note that the temperature signals are expressed in $^{\circ}C$, whereas precipitation signals are expressed as percentages.

dataset are missing. Despite this partial data availability, the model retains the same architecture as the one trained without missing points, and thus can still generate predictions across the entire predictand domain.

In Figure 8, we present the Root Mean Squared Error (RMSE) on the test set for the three downscaled variables under the different training regimes. Each regime is trained on versions of the ~~ROCIO-IBEB~~ STATIONS-IBEB dataset with varying percentages of artificially introduced missing values, shown on the x-axis of each subplot. Note that the RMSE is computed on the test partition without missing data, enabling us to evaluate the model~~'~~'s ability to extrapolate to data points that were partially missing during training. Overall, ~~fine-tuning leads to better performance in the presence of missing data in~~ the ~~STATIONS-IBEB dataset~~transferred relationships from the pre-trained model prove beneficial in low-data regimes, as evidenced by the improved performance of fine-tuned models compared to training from scratch. For both minimum and maximum temperatures, the fine-tuned regimes achieve lower RMSE values than the fully-trained regime, converging to the same value when the training is carried out on the STATIONS-IBEB dataset without missing data. The advantage of ~~fine-tuning~~ transferring pre-learned representations is especially pronounced under high missing-data regimes (e.g., 90%), where ~~having pre-learned representations aids~~ these relationships aid in generalizing to unseen data points. For precipitation, a similar trend is observed between the fully-training and both fine-tuning regimes~~;~~, however, the fully fine-tuned model performs worse than

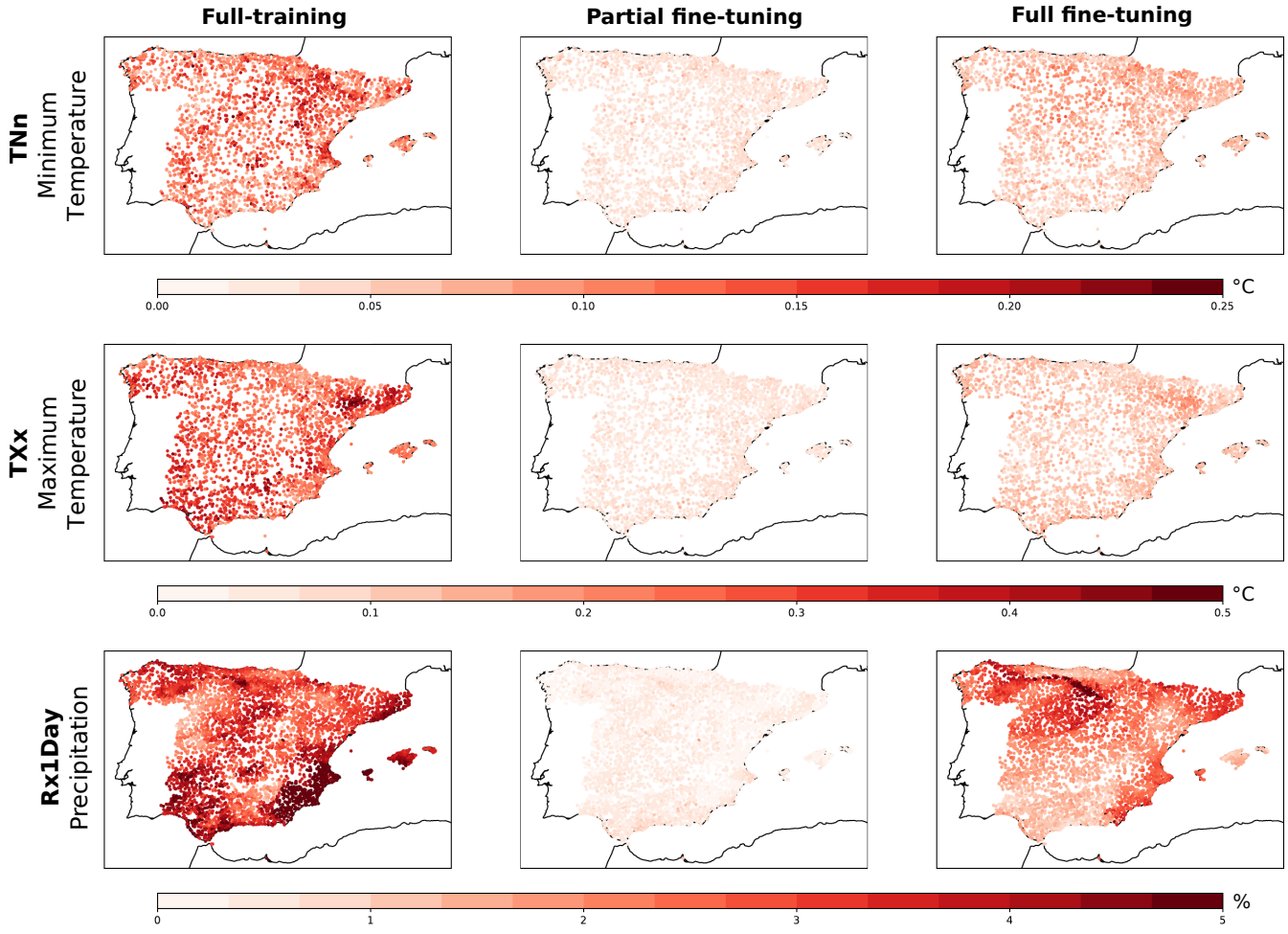

Figure 7. Standard deviation of the climate change signals for the TNn, TXx, and RX1Day indices (shown in rows), computed across ten training replicas under the fully-training, partial fine-tuning and full fine-tuning regimes on STATIONS-IBEB. The standard deviation for temperature signals is expressed in $^{\circ}C$, while that for precipitation signals is given as a percentage.

the others when more than 60% of the training data is missing, though it still outperforms the fully-trained model beyond this threshold.

4.5 Application to a Regional Dataset

Finally, we extend the analysis of the proposed methodology to the STATIONS-CAT dataset. As discussed in Section 2.1.1, this dataset exemplifies a common use case motivating this approach: a regional, station-based observational dataset with limited temporal coverage. In contrast to STATIONS-IBEB, which is partially used to construct the ROCIO-IBEB gridded dataset, STATIONS-CAT comprises an independent set of stations over Catalonia and has no relation with ROCIO-IBEB.

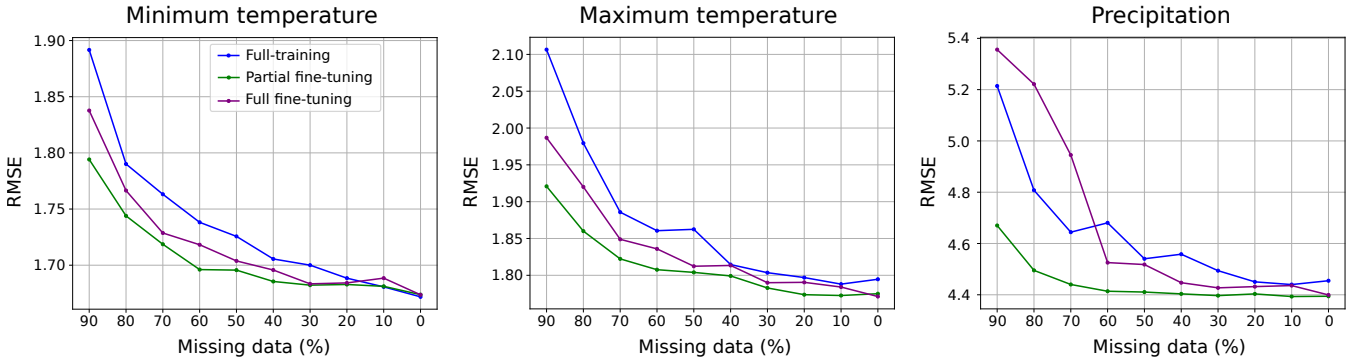

Figure 8. Root Mean Squared Error (RMSE) computed over the test set for three different training regimes (fully-training, partial fine-tuning and full fine-tuning), shown for the three downscaled variables (minimum temperature, maximum temperature and precipitation), in columns. Each model is trained on a different version of the ~~ROCIO-IBEB~~ STATIONS-IBEB dataset, varying in the proportion of artificially introduced missing data (indicated on the x-axis of each subplot). Note that the RMSE shown here is computed using the ~~ROCIO-IBEB~~ STATIONS-IBEB test partition without missing data.

For STATIONS-CAT, data availability for all stations begins in 2009. Consequently, the fine-tuning regimes are trained using data from only the 2009–2019 period, while the test set covers 2019–2021. This experimental setup reflects one of the main challenges in regional statistical downscaling, namely the scarcity of long, homogeneous observational records. For the sake of clarity, the analysis in this section focuses on precipitation, which is generally considered the most challenging variable to downscale (Rampal et al., 2024; González-Abad, 2025a). The pre-trained model based on the ROCIO-IBEB dataset is the same as that used in the previous analysis with STATIONS-IBEB, and the training configurations for the full-training, partial fine-tuning and full fine-tuning regimes remain unchanged.

As in Section 4.2, Figure 9 shows the ASM for the pre-training model (on ROCIO-IBEB) and the full-training, partial fine-tuning, and full fine-tuning regimes for STATIONS-CAT (for precipitation downscaling). In this case, the ASMs for both datasets are computed over the period corresponding to the STATIONS-CAT test set (2019-2021). It is important to note that, for the pre-training model on ROCIO-IBEB, the ASM is computed using the same number of grid points as there are stations in STATIONS-CAT, selecting the grid points closest to each corresponding station. This ensures a fair comparison between the relationships exposed by these techniques. The ASM for STATIONS-CAT exhibits a pattern similar to that observed for STATIONS-IBEB, with relevance highly distributed across the different predictor variables. It is also noticeable that, despite covering different regions, the distribution of relevance across variables for the pre-trained model is broadly consistent with that shown in Figure 5. Focusing again in Figure 9, the ASMs for the full-training and the partial and full fine-tuning regimes display large discrepancies, indicating that models trained exclusively on STATIONS-CAT tend to learn relationships that are less aligned with the physical patterns encoded in the reference pre-trained model on ROCIO-IBEB.

To illustrate the spatial distribution of relevance (i.e., the regions to which the model attends when making predictions for a specific station) we compute the saliency maps for the pre-training model (on ROCIO-IBEB) and the full-training, partial

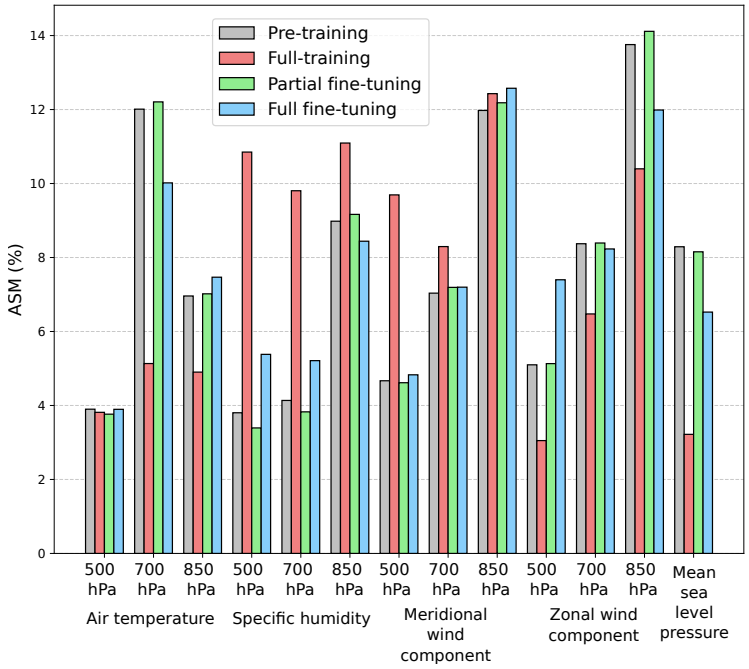

Figure 9. Same as Figure 5, but for the full-training, partial fine-tuning, and full fine-tuning precipitation downscaling models applied to the STATIONS-CAT dataset over the test period 2019–2021. For the pre-trained model (on ROCIO-IBEB), the ASM is computed using the same number of grid points as there are stations in STATIONS-CAT, selecting the grid points closest to each corresponding station.

fine-tuning, and full fine-tuning models (on STATIONS-IBEB) for a single station corresponding to Sabadell Airport. For the pre-training model, the saliency map is computed for the grid point closest to this station. Saliency maps are calculated for each day of the 2019–2021 period (the STATIONS-CAT test set), and then averaged to obtain the mean saliency for precipitation downscaling at this station. Figure 10 presents these average saliency maps for the four training regimes (shown in columns) and for four predictor variables (air temperature at 850 hPa, specific humidity at 700 hPa, meridional wind component at 850 hPa, and mean sea level pressure), shown in rows.

Overall, the saliency maps in Figure 5 show that, for all models, the learned relationships are predominantly local, focused around the station being analyzed. This is expected for daily downscaling, as the physical processes governing local-scale precipitation dynamics generally occur within nearby regions. Comparing the spatial distribution of relevance across models reveals several interesting patterns. The saliency maps for the pre-training and partial fine-tuning models are highly similar, consistent with the discussion in Section 3: the important relationships are primarily learned by the feature extractor, while the final layer acts mainly as a linear calibrator. In contrast, the full fine-tuning model, in which the feature extractor weights are also updated starting from the pre-trained model, shows some changes in the spatial distribution of relevance. Nevertheless, the expected physical relationships are still captured, for example, the dependence of precipitation in Catalonia on conditions over the Mediterranean. By comparison, the full-training model, trained solely on STATIONS-CAT, exhibits a different spatial

distribution of relevance. It places more attention on regions in the central Iberian Peninsula, as seen, for example, in the saliency maps of specific humidity at 700 hPa and meridional wind at 850 hPa. This suggests that the relationships learned by the full-training model may deviate from the expected physical dynamics governing precipitation in the region, highlighting the value of pre-training in constraining the model toward physically plausible relationships.

Finally, in Figure 11 we present the same analysis as in Figure 7 (the standard deviation of the climate-change signal across training replicas) focusing specifically on the RX1Day index for precipitation. Similar to the results for STATIONS-IBEB, both fine-tuning regimes reduce the variability of the climate change signal across replicas, with partial fine-tuning achieving notably low variation. The full fine-tuning regime shows intermediate variability, higher than the partial fine-tuning but lower than full-training, with elevated variability particularly in the northwestern region corresponding to the Pyrenees Mountains.

5 Discussion

The results presented in this work demonstrate the potential of pre-training for PP downscaling, particularly for developing a standard model in the region of Spain. ~~Specifically, as~~ As clarified in the introduction, the primary objective of this approach is not to maximize predictive skill on individual datasets, but rather to ensure consistency with a standard reference model while enhancing the robustness of climate projections across different regional datasets.

A key contribution of this work lies in demonstrating how fine-tuning facilitates the transfer of learned representations from the pre-trained model to models trained on regional datasets. As shown in Section 4.1, ~~fine-tuned DeepESD models show accelerated convergence, indicating that representations learned on the ROCIO-IBEB dataset contain valuable information for modeling the point-based dataset~~the performance metrics for STATIONS-IBEB across the three training regimes (full-training, partial fine-tuning, and full fine-tuning) are comparable. However, this ~~benefit does not necessarily translate into improved accuracy on STATIONS-IBEB, likely due to the presence of higher and more localized extreme values, which are more challenging to model than their smoothed counterparts in the interpolated ROCIO-IBEB gridded dataset . In the case of precipitation, fine-tuning the feature extractor appears to be beneficial~~similarity in skill masks a more fundamental difference: the fine-tuned models inherit the physical relationships encoded in the pre-trained model, whereas the fully-trained model learns its own set of relationships that may diverge from the established baseline. This inheritance of learned representations is particularly valuable when working with limited or regional datasets, where models trained from scratch are more susceptible to overfitting and learning spurious patterns.

The application to the STATIONS-CAT dataset provides evidence for the value of this approach in realistic operational scenarios. This regional dataset, covering only Catalonia with limited temporal coverage (2009–2021), exemplifies the typical challenges faced when developing regional climate projections: fewer stations and shorter time periods. Under these conditions, models trained exclusively on STATIONS-CAT show substantial deviations in their learned relationships, as evidenced by both the ASM diagnostics (Figure 9) and the saliency maps (Figure 10). The ASM for the full-training regime exhibits large discrepancies compared to the pre-trained model, indicating that ~~for this variable, slight adjustments to the learned large-scale patterns may be necessary to better capture local characteristics~~with limited data availability, models may fit noise and learn

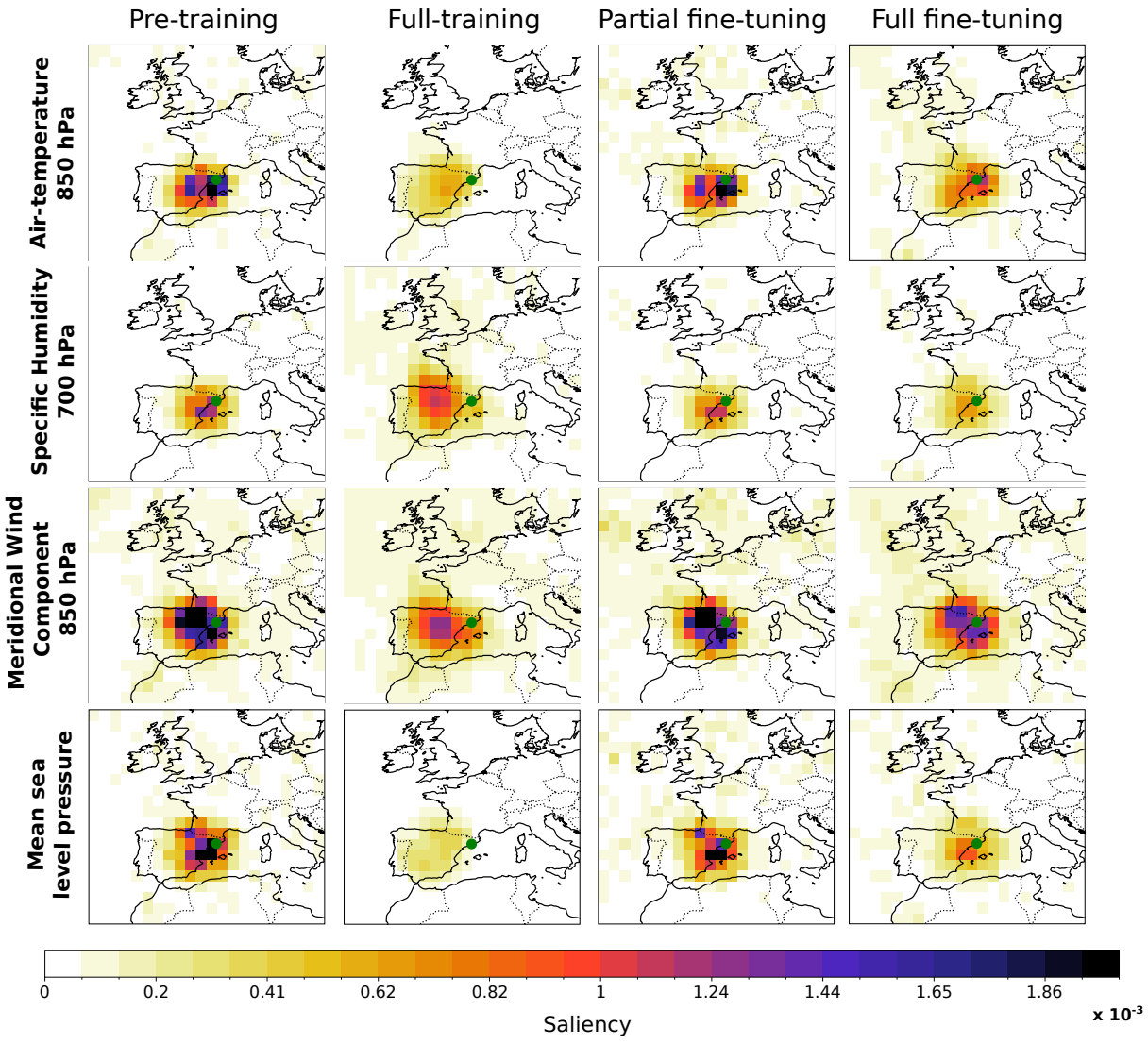

Figure 10. Saliency maps for the different training regimes: pre-training on ROCIO-IBEB, and full training, partial fine-tuning, and full fine-tuning on STATIONS-CAT (shown in columns), computed and aggregated over the STATIONS-CAT test period (2019–2021). Saliency maps are originally computed over all predictor variables but for clarity, we display only a subset of variables and a spatial domain centered on Catalonia. The variables shown (in columns) are air temperature at 850 hPa, specific humidity at 700 hPa, meridional wind component at 850 hPa, and mean sea level pressure. Saliency maps are computed with respect to the station indicated by the green point, corresponding to the Sabadell airport. For the pre-trained model the saliency maps are computed for the grid point closest to this station.

relationships that are less aligned with known dynamics. In contrast, both fine-tuning strategies successfully transfer the representations learned on ROCIO-IBEB, maintaining physical consistency even when adapted to this independent regional dataset.

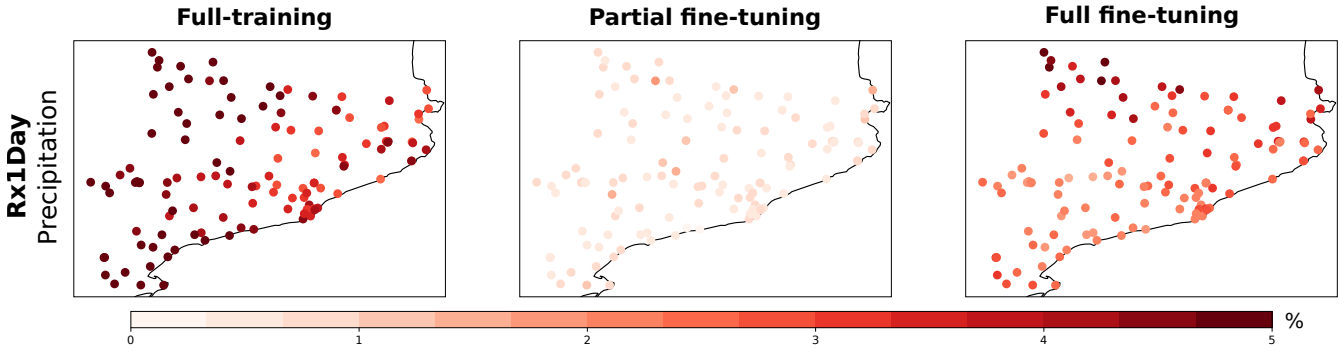

Figure 11. Same as Figure 7 but for the fully-training, partial fine-tuning and full fine-tuning regimes on STATIONS-CAT dataset.

When analyzing the patterns learned by the various models ~~, by focusing on~~ using the ASM diagnostic, we find differences in the distribution of relevance across predictors for the three downscaled variables. Focusing on minimum temperature, all models exhibit their highest relevance in air temperature and specific humidity at 850 hPa (the lowest level) and mean sea level pressure, although the fully-trained model also shows a relatively high relevance for the zonal wind component at 850 hPa. This finding aligns with previous studies ~~(Baño-Medina, 2021; González-Abad et al., 2023)~~(Baño-Medina et al., 2023; González-Abad et al., 2023) , which indicate that, for temperature, the most critical variables are air temperature and specific humidity at 1000 hPa, alongside geopotential height at 1000 hPa—closely related to mean sea-level pressure in our setup. The fully-trained and fully fine-tuned models show the greatest deviation from the expected relevance distribution, particularly in the increased ASM assigned to mean sea level pressure by the fully fine-tuned model. Nonetheless, the overall relevance patterns remain broadly similar. In contrast, the partial fine-tuned model exhibits relevance distributions more closely aligned with those of the pre-trained model, suggesting that, for this variable, fine-tuning the feature extractor may introduce a divergence from the relationships learned by the pre-trained model.

The ASM diagnostic indicates a stronger relationship between mean sea level pressure and maximum temperature than with minimum temperature. This is consistent with our understanding of temperature variability. Maximum temperature is largely controlled by atmospheric circulation patterns and adiabatic warming, both of which are effectively captured by this predictor. In contrast, minimum temperature typically requires a more complex set of predictors due to the influence of local surface conditions, moisture availability and radiative cooling processes. This distinction aligns with the fundamental meteorological principles governing daytime heating and nighttime cooling, as well as with previous studies (Favà et al., 2016, 2018; Merino et al., 2018; Pérez and García, 2023) that have demonstrated a robust relationship between pressure-related variables and temperature variability over the Iberian Peninsula. Interestingly, recent work (Cariou et al., 2025) has shown that DL models can effectively reproduce daily temperature variations over Europe using only mean sea level pressure as input, thereby underscoring the strong influence of atmospheric circulation. This aligns with our XAI-based findings, which highlight the particular relevance of mean sea level pressure, especially for maximum temperature.

In the case of precipitation, relevance is distributed across the full set of predictor variables, which aligns with the fact that precipitation dynamics encompass a broader range of phenomena. As shown in ~~Figure 5~~ Figures 5 and 9, wind components play a significant role, particularly in northwestern regions, where westerly winds transport moisture from the Atlantic Ocean. The influence of both temperature and humidity also makes sense in this context; for example, higher temperatures require greater amounts of humidity to reach saturation, a key factor in precipitation events. Additionally, mean sea-level pressure is relevant because pressure differentials can trigger the lifting of moist air. Overall, while all training regimes agree on the distribution of relevance for STATIONS-IBEB, the partially fine-tuned model produces values that more closely align with those of the pre-trained model. For STATIONS-CAT, this alignment becomes even more critical, as the full-training model shows substantial deviations that suggest the learning of spurious relationships.

The spatial saliency maps presented in Figure 10 provide further insight into the nature of these learned relationships. Overall, these maps show that for all models, the learned relationships are predominantly local, focused around the station being analyzed. This is expected for daily downscaling, as the physical processes governing local-scale precipitation dynamics generally occur within nearby regions, consistent with findings in previous work (González-Abad et al., 2023). The spatial saliency maps reveal that while both fine-tuning strategies maintain similarity to the pre-trained model's patterns, the full-training model exhibits notably different spatial distributions of relevance, placing more attention to the central Iberian Peninsula rather than the Mediterranean region expected for precipitation in Catalonia. This deviation highlights how training exclusively on limited regional data can lead to relationships that diverge from established dynamics, underscoring the value of pre-training in constraining models toward physically plausible patterns.

Regarding the climate change signals~~produced by these models~~ , the results for minimum and maximum temperature ~~, the results~~ closely align with those reported in ~~González-Abad and Gutiérrez (2025)~~(González-Abad and Gutiérrez, 2025), indicating plausible projections. The lack of relevant differences among the signals from the different training regimes for STATIONS-IBEB may originate from the use of 30-year climatologies, suggesting that the choice between ~~the ROCIO-IBEB and STATIONS-IBEB~~ datasets does not fundamentally impact statistical trends at these temporal scales ~~.~~ when sufficient data is available. However, the enhanced robustness provided by fine-tuning becomes evident when examining variability across training replicas. As expected, the partially fine-tuned model ~~exhibit~~ exhibits lower variability than the fully fine-tuned ~~one since fixing a subset of parameters~~ model, since fixing the feature extractor constrains the optimization process ~~, limiting it~~ to a smaller region of the parameter space ~~and leading to convergence within similar local minima (Erhan et al., 2010). Overall, if the pre-trained model is deemed reliable,~~ (Erhan et al., 2010). This is evident in Figures 7 and 11, where the spread across replicas becomes nearly zero when the feature extractor is frozen. For STATIONS-CAT, this reduction in variability takes on particular significance: both fine-tuning ~~strategies can help reduce epistemic uncertainty, reinforcing confidence in the projections~~regimes substantially reduce the variability of the climate change signal compared to full-training, with partial fine-tuning achieving notably low variation. This enhanced robustness can be attributed to the transfer of meaningful physical relationships from the pre-trained model, making fine-tuned models less susceptible to the spurious patterns that emerge when training exclusively on limited regional data, a reduction in epistemic uncertainty particularly valuable for developing reliable regional climate projections that maintain consistency with national-scale assessments.

As shown in Section 4.4, ~~pre-training is beneficial~~ the benefits of transferring pre-learned representations are also evident in low-data regimes, ~~as it provides new DL models with useful representations~~where portions of the observational record are missing. The transferred relationships from the pre-trained model prove beneficial when data availability is limited, as evidenced by the improved performance of fine-tuned models compared to training from scratch. This advantage has been previously observed in the ~~SR~~ super-resolution field (Zhu and Zhou, 2024), where pre-trained models have successfully transferred relationships across different regions. However, in ~~SR~~super-resolution, regional transferability is more straightforward since the model learns an interpolation function by relying on a coarser version of the predictand. In contrast, in the PP approach, the empirical relationships being learned may depend on dynamic regional phenomena, making ~~them potentially less transferable across different regions~~the successful transfer demonstrated here particularly noteworthy.

6 Conclusions

Current statistical downscaling methodologies often involve different research groups employing various schemes and drawing on datasets of varying spatial resolution and temporal coverage; this can lead to inconsistencies in the resulting downscaled projections. Such discrepancies pose challenges when nationwide products exist—for example, Escenarios-PNACC 2024 in Spain~~—as discrepancies may arise~~. To address this issue, we introduce pre-training as a strategy to develop a standard DL model for PP downscaling in Spain, ~~ensuring consistency when adapting it~~ with the explicit goal of ensuring consistency and enhancing robustness when adapting to different regional datasets. Specifically, we propose a standard model ~~by using as the basis the~~ using the DeepESD architecture trained on the ROCIO-IBEB gridded dataset as the foundation. We then demonstrate how this model can serve as a ~~foundation~~ basis for training on ~~a~~ point-based ~~dataset (STATIONS-IBEB) using two~~ datasets through two fine-tuning approaches: partial ~~and full fine-tuning~~(freezing the feature extractor) and full (allowing the feature extractor to adapt).

As demonstrated by our XAI-based analysis, the pre-training paradigm enables ~~the resulting model~~ fine-tuned models to inherit the physical relationships learned by the pre-trained model, grounding ~~its~~ their climate projections in ~~the same~~ a consistent physical basis. In the case of Spain, these inherited relationships align with previous findings in the literature ~~(Baño-Medina, 2021; González-Abad et al., 2023)~~(Baño-Medina et al., 2021; González-Abad et al., 2023), and the fine-tuned versions ~~follow suit. Such~~ successfully preserve them even when adapted to independent regional datasets. This inheritance is particularly ~~promising when transferring to~~ valuable when working with small regional datasets, where DL models trained from scratch are prone to overfitting and learning spurious patterns~~(González-Abad et al., 2023). Moreover~~, as evidenced by our results with STATIONS-CAT. The application to this regional Catalonian dataset exemplifies the typical operational scenario motivating this work: developing regional climate projections that maintain coherence with established national products while accommodating local observational data constraints.

Beyond ensuring consistency, leveraging an established downscaling model through fine-tuning reduces the epistemic uncertainty ~~of the fine-tuned model, as the~~ in climate projections. By constraining the parameter search space ~~is constrained to that of the model taken as the basis. When this~~ and transferring meaningful representations, fine-tuning strategies lead to

more robust projections with reduced variability across model initializations. This is particularly evident in the climate change signals for both STATIONS-IBEB and STATIONS-CAT, where fine-tuning substantially decreases the spread across training replicas compared to training from scratch. When the pre-trained model is proven ~~to be~~ reliable, this reduction in uncertainty further enhances trust in ~~the final projections . Beyond ensuring greater consistency with this model, our work suggests that~~ regional projections and ensures their alignment with national assessments.

Our work also reveals additional benefits of pre-training ~~offers additional benefits~~ in the context of PP downscaling. ~~For instance, if the datasets used for the pre-trained and the fine-tuned models share certain characteristics (e.g., a similar geographic region), the fine-tuned model converges faster by effectively leveraging the representations learned by the pre-trained model. Furthermore, in~~ In low-data regimes—where datasets often have missing observations due to station errors or ~~other issues~~limited temporal coverage—fine-tuning allows ~~the DL model~~ DL models to rely on ~~the already learned relationships~~ pre-learned representations rather than attempting to extract all relationships from sparse local data. This is especially valuable in regional contexts, where incomplete or inconsistent datasets are common.

The promising results of pre-training in the context of PP downscaling open several avenues for future research. In the case of Spain, one potential approach to creating a more representative basis model is to leverage the range of available observational datasets ~~(Peral García et al., 2017; Herrera et al., 2019)~~(Peral García et al., 2017; Herrera et al., 2019), which are generated using various methods (e.g., different interpolation techniques). Integrating multiple datasets could help reduce observational uncertainty tied to dataset selection. Another promising direction is to extend this approach to larger regions, such as Europe, where multiple observational datasets are also available. In such continental regions, downscaling models are typically trained on continental or global observational datasets that may not reflect the specific properties of individual regions~~(see, e.g., Baño-Medina et al. (2022))~~. In this case, a standard model could be trained on broader datasets and then ~~transferred~~ fine-tuned with regional datasets at higher resolutions, further enhancing its accuracy and applicability in generating nationwide products while maintaining consistency across scales. Lastly, following efforts both within (Lessig et al., 2023; Nguyen et al., 2023; Bodnar et al., 2024) and beyond (Bommasani et al., 2021) the climate and weather domains, a self-supervised learning strategy could be employed to train a foundational model by drawing on data from GCMs and Regional Climate Models (RCMs) (Rampal et al., 2024). Such a foundational model could then serve as a basis for numerous regions, variables, and even different downscaling tasks (e.g., PP downscaling or RCM emulation).

Overall, pre-training emerges as a ~~promising~~ valuable strategy for developing ~~a standard PP downscaling model, particularly for Spain, the focus of this study~~consistent and robust regional climate projections. By fine-tuning a reliable pre-trained model, we can ~~retain its physical basis, ensuring consistent projections and reducing~~ ensure that regional projections inherit established physical relationships, maintain coherence with national-scale assessments, and exhibit reduced epistemic uncertainty in critical extrapolation scenarios, such as downscaling GCMs under future climate conditions. ~~Moreover, pre-training confers additional benefits, including enhanced generalization in low-data regimes~~The successful application to STATIONS-CAT demonstrates that this approach is particularly beneficial for the typical operational scenario: adapting a standard national model to limited regional datasets while preserving physically meaningful relationships and enhancing projection robustness. These

advantages position pre-training as a valuable avenue for future research, potentially enabling robust ~~pre-trained~~ standard models that a wider community of users and researchers can readily adopt for developing consistent multi-scale climate projections.

Code and data availability. All the code and data required to reproduce the experiments in this study are publicly available. The processed ERA5 dataset and the EC-Earth3-Veg data are freely accessible via Zenodo (https://zenodo.org/records/16687087, González-Abad 2025b), while the ROCIO-IBEB and STATIONS-IBEB datasets are also available on Zenodo (https://zenodo.org/records/17338349, González-Abad and AEMET). The daily blended station data from the European Climate Assessment & Dataset (ECA&D) can be obtained from https://www.ecad.eu/dailydata/predefinedseries.php. The code to fully reproduce the experiments can be found at https://zenodo.org/records/18467873, (González-Abad 2026). In addition, all trained models used to generate the results are provided in a dedicated Zenodo repository (https://zenodo.org/records/18468086, González-Abad 2025a), ensuring full reproducibility of the findings presented in this manuscript.

Author contributions. The conceptualization of the study was carried out by JGA and JMG. Code development was performed by JGA, while data acquisition and processing were conducted by JGA, MI and AH. Formal analysis was undertaken by JGA and JMG, and visualization was prepared by JGA. All authors contributed to the writing, reviewing, and editing of the manuscript.

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

Acknowledgements. We would like to acknowledge all the teams involved in the production and maintenance of the ERA5 reanalysis dataset and the EC-Earth3 climate model simulations. We express our special gratitude to the teams at AEMET for the development and provision of the ROCIO-IBEB and STATIONS-IBEB datasets~~used in this work~~, as well as the European Climate Assessment & Dataset (ECA&D) project for the STATIONS-CAT dataset. We also thank Sixto Herrera García for suggesting us the use of ECA&D dataset as well as for his support in downloading and processing the data. González-Abad acknowledges support from grant CPP2021-008510 funded by MICIU/AEI/10.13039/501100011033 and by the "European Union" and the "European Union NextGenerationEU/PRTR". This research work was supported by the Ministry for the Ecological Transition and the Demographic Challenge (MITECO) and the European Commission NextGenerationEU (Regulation EU 2020/2094), through CSIC's Interdisciplinary Thematic Platform Clima (PTI-Clima).

References

Adadi, A. and Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), IEEE access, 6, 52 138–52 160, 2018.

Addison, H., Kendon, E., Ravuri, S., Aitchison, L., and Watson, P. A.: Machine learning emulation of precipitation from km-scale regional climate simulations using a diffusion model, arXiv preprint arXiv:2407.14158, 2024.

Agrawal, P., Girshick, R., and Malik, J.: Analyzing the performance of multilayer neural networks for object recognition, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13, pp. 329–344, Springer, 2014.

Amblar-Francés, M. P., Ramos-Calzado, P., Sanchis-Lladó, J., Hernanz-Lázaro, A., Peral-García, M. C., Navascués, B., Dominguez-Alonso, M., Pastor-Saavedra, M. A., and Rodríguez-Camino, E.: High resolution climate change projections for the Pyrenees region, Advances in Science and Research, 17, 191–208, 2020.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al.: Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information fusion, 58, 82–115, 2020.

Balmaceda-Huarte, R., Baño-Medina, J., Olmo, M. E., and Bettolli, M. L.: On the use of convolutional neural networks for downscaling daily temperatures over southern South America in a climate change scenario, Climate Dynamics, 62, 383–397, 2024.

Baño-Medina, J., Manzanas, R., and Gutiérrez, J. M.: Configuration and intercomparison of deep learning neural models for statistical downscaling, Geoscientific Model Development, 13, 2109–2124, 2020.

Baño-Medina, J., Manzanas, R., and Gutiérrez, J. M.: On the suitability of deep convolutional neural networks for continental-wide downscaling of climate change projections, Climate Dynamics, 57, 2941–2951, 2021.

Baño-Medina, J., Manzanas, R., Cimadevilla, E., Fernández, J., González-Abad, J., Cofiño, A. S., and Gutiérrez, J. M.: Downscaling multi-model climate projection ensembles with deep learning (DeepESD): contribution to CORDEX EUR-44, Geoscientific Model Development Discussions, 2022, 1–14, 2022.

Basque Government: Klimatek: Downscaled Climate Change Scenarios for the Basque Autonomous Community, Technical report, Basque Government – Ihobe, available online: https://www.euskadi.eus/contenidos/documentacion/escenarios_cc/es_def/adjuntos/Klimatek-2020.pdf, 2020.

Baño-Medina, J.: Understanding Deep Learning Decisions in Statistical Downscaling Models, in: Proceedings of the 10th International Conference on Climate Informatics, CI2020, pp. 79–85, Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3429309.3429321, 2021.

Baño-Medina, J., Iturbide, M., Fernández, J., and Gutiérrez, J. M.: Transferability and explainability of deep learning emulators for regional climate model projections: Perspectives for future applications, arXiv preprint arXiv:2311.03378, 2023.

Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H.: Greedy layer-wise training of deep networks, Advances in neural information processing systems, 19, 2006.

Bishop, C. M. and Nasrabadi, N. M.: Pattern recognition and machine learning, Springer, 2006.

Bodnar, C., Bruinsma, W. P., Lucic, A., Stanley, M., Brandstetter, J., Garvan, P., Riechert, M., Weyn, J., Dong, H., Vaughan, A., et al.: Aurora: A foundation model of the atmosphere, arXiv preprint arXiv:2405.13063, 2024.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models, arXiv preprint arXiv:2108.07258, 2021.

Brown, T. B.: Language models are few-shot learners, arXiv preprint arXiv:2005.14165, 2020.

Cariou, E., Cattiaux, J., Qasmi, S., Ribes, A., Cassou, C., and Doury, A.: Linking European temperature variations to atmospheric circulation with a neural network: A pilot study in a climate model, Geophysical Research Letters, 52, e2024GL113 540, 2025.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A.: Emerging properties in self-supervised vision transformers, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 9650–9660, 2021.

Chen, D., Rojas, M., Samset, B., Cobb, K., Diongue Niang, A., Edwards, P., Emori, S., Faria, S., Hawkins, E., Hope, P., Huybrechts, P., Meinshausen, M., Mustafa, S., Plattner, G.-K., and Tréguier, A.-M.: Framing, Context, and Methods, in: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, edited by Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J., Maycock, T., Waterfield, T., Yelekçi, O., Yu, R., and Zhou, B., pp. 147–286, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, https://www.ipcc.ch/report/ar6/wg1/chapter/chapter-1, 2021.

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., and Sutskever, I.: Generative pretraining from pixels, in: International conference on machine learning, pp. 1691–1703, PMLR, 2020.

Cornes, R. C., van der Schrier, G., van den Besselaar, E. J., and Jones, P. D.: An ensemble version of the E-OBS temperature and precipitation data sets, Journal of Geophysical Research: Atmospheres, 123, 9391–9409, 2018.

Correa, C., Hernanz Lázaro, A., and Rodríguez Guisado, E.: Evaluación de métodos de regionalización estadística para la generación de proyecciones climáticas en el marco del PNACC-2 2021-2030, 2023.

Cos, J., Doblas-Reyes, F., Jury, M., Marcos, R., Bretonnière, P.-A., and Samsó, M.: The Mediterranean climate change hotspot in the CMIP5 and CMIP6 projections, Earth System Dynamics, 13, 321–340, 2022.

Devlin, J.: Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805, 2018.

Döscher, R., Acosta, M., Alessandri, A., Anthoni, P., Arneth, A., Arsouze, T., Bergmann, T., Bernadello, R., Bousetta, S., Caron, L.-P., et al.: The EC-earth3 Earth system model for the climate model intercomparison project 6, Geoscientific Model Development Discussions, 2021, 1–90, 2021.

Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929, 2020.

Doury, A., Somot, S., Gadat, S., Ribes, A., and Corre, L.: Regional climate model emulator based on deep learning: Concept and first evaluation of a novel hybrid downscaling approach, Climate Dynamics, 60, 1751–1779, 2023.

Doury, A., Somot, S., and Gadat, S.: On the suitability of a convolutional neural network based RCM-emulator for fine spatio-temporal precipitation, Climate Dynamics, 62, 8587–8613, 2024.

Erhan, D., Courville, A., Bengio, Y., and Vincent, P.: Why does unsupervised pre-training help deep learning?, in: Proceedings of the thirteenth international conference on artificial intelligence and statistics, pp. 201–208, JMLR Workshop and Conference Proceedings, 2010.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geoscientific Model Development, 9, 1937–1958, 2016.

Favà, V., Curto, J., and Llasat, M.: Regional differential behaviour of maximum temperatures in the Iberian Peninsula regarding the Summer NAO in the second half of the twentieth century, Atmospheric Research, 182, 319–334, 2016.

Favà, V., Curto, J. J., and Llasat, M. d. C.: Changes in summer pressure patterns across the late 1960s and their influence on temperature trends on the eastern coast of the Iberian Peninsula, Atmosphere, 9, 42, 2018.

François, B., Thao, S., and Vrac, M.: Adjusting spatial dependence of climate model outputs with cycle-consistent adversarial networks, Climate dynamics, 57, 3323–3353, 2021.

Glorot, X. and Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the thirteenth international conference on artificial intelligence and statistics, pp. 249–256, JMLR Workshop and Conference Proceedings, 2010.

González-Abad, J. and Gutiérrez, J. M.: Are deep learning methods suitable for downscaling global climate projections? An intercomparison for temperature and precipitation over Spain, Artificial Intelligence for the Earth Systems, 4, 240 121, 2025.

González-Abad, J., Baño-Medina, J., and Gutiérrez, J. M.: Using explainability to inform statistical downscaling based on deep learning beyond standard validation approaches, Journal of Advances in Modeling Earth Systems, 15, e2023MS003 641, 2023.

González-Abad, J.: Deep learning models used in the manuscript "Pre-training for Deep Statistical Climate Downscaling: A case study within the Spanish National Adaptation Plan (PNACC)", https://doi.org/10.5281/zenodo.18468086, 2025a.

González-Abad, J.: Predictors for the manuscript "Pre-training for Deep Statistical Climate Downscaling: A case study within the Spanish National Adaptation Plan (PNACC)"", https://doi.org/10.5281/zenodo.16687087, 2025b.

González-Abad, J.: jgonzalezab/pretraining_PP_downscaling: v1.0, https://doi.org/10.5281/zenodo.18467873, 2026.

González-Abad, J. and (AEMET), S. M. A.: Predictands for the manuscript "Pre-training for Deep Statistical Climate Downscaling: A case study within the Spanish National Adaptation Plan (PNACC)", https://doi.org/10.5281/zenodo.17338348, 2025.

Goodfellow, I., Bengio, Y., and Courville, A.: Deep learning, MIT press, 2016.

Gutiérrez, J. M., San-Martín, D., Brands, S., Manzanas, R., and Herrera, S.: Reassessing statistical downscaling techniques for their robust application under climate change conditions, Journal of Climate, 26, 171–188, 2013.

Gutiérrez, J. M., Maraun, D., Widmann, M., Huth, R., Hertig, E., Benestad, R., Rössler, O., Wibig, J., Wilcke, R., Kotlarski, S., et al.: An intercomparison of a large ensemble of statistical downscaling methods over Europe: Results from the VALUE perfect predictor cross-validation experiment, International journal of climatology, 39, 3750–3785, 2019.

Hastie, T., Friedman, J., and Tibshirani, R.: The Elements of Statistical Learning, Springer Series in Statistics, Springer, New York, NY, http://link.springer.com/10.1007/978-0-387-21606-5, 2001.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R.: Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 16 000–16 009, 2022.

Hernanz, A., García-Valero, J. A., Domínguez, M., Ramos-Calzado, P., Pastor-Saavedra, M. A., and Rodríguez-Camino, E.: Evaluation of statistical downscaling methods for climate change projections over Spain: present conditions with perfect predictors, International Journal of Climatology, 42, 762–776, 2022.

Herrera, S., Cardoso, R. M., Soares, P. M., Espírito-Santo, F., Viterbo, P., and Gutiérrez, J. M.: Iberia01: A new gridded dataset of daily precipitation and temperatures over Iberia, Earth System Science Data, 11, 1947–1956, 2019.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, 2020.

Hoerling, M., Eischeid, J., Perlwitz, J., Quan, X., Zhang, T., and Pegion, P.: On the increased frequency of Mediterranean drought, Journal of climate, 25, 2146–2161, 2012.

Karger, D. N., Nobis, M. P., Normand, S., Graham, C. H., and Zimmermann, N. E.: CHELSA-TraCE21k–high-resolution (1 km) downscaled transient temperature and precipitation data since the Last Glacial Maximum, Climate of the Past, 19, 439–456, 2023.

Kenton, J. D. M.-W. C. and Toutanova, L. K.: Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of naacL-HLT, vol. 1, p. 2, Minneapolis, Minnesota, 2019.

Kheir, A. M., Elnashar, A., Mosad, A., and Govind, A.: An improved deep learning procedure for statistical downscaling of climate data, Heliyon, 9, 2023.

Kingma, D. P.: Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980, 2014.

Klein Tank, A. M., Wijngaard, J., Können, G., Böhm, R., Demarée, G., Gocheva, A., Mileta, M., Pashiardis, S., Hejkrlik, L., Kern-Hansen, C., et al.: Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment, International Journal of Climatology: A Journal of the Royal Meteorological Society, 22, 1441–1453, 2002.

Klok, E. and Klein Tank, A.: Updated and extended European dataset of daily climate observations, International Journal of Climatology: A Journal of the Royal Meteorological Society, 29, 1182–1191, 2009.

Krizhevsky, A., Sutskever, I., and Hinton, G. E.: Imagenet classification with deep convolutional neural networks, Advances in neural information processing systems, 25, 2012.

LeCun, Y., Bengio, Y., et al.: Convolutional networks for images, speech, and time series, The handbook of brain theory and neural networks, 3361, 1995, 1995.

Lessig, C., Luise, I., Gong, B., Langguth, M., Stadtler, S., and Schultz, M.: AtmoRep: A stochastic model of atmosphere dynamics using large scale representation learning, arXiv preprint arXiv:2308.13280, 2023.

Maraun, D. and Widmann, M.: Statistical downscaling and bias correction for climate research, Cambridge University Press, 2018.

Merino, A., Martín, M., Fernández-González, S., Sánchez, J., and Valero, F.: Extreme maximum temperature events and their relationships with large-scale modes: potential hazard on the Iberian Peninsula, Theoretical and applied climatology, 133, 531–550, 2018.

Minh, D., Wang, H. X., Li, Y. F., and Nguyen, T. N.: Explainable artificial intelligence: a comprehensive review, Artificial Intelligence Review, pp. 1–66, 2022.

Miró, J. J., Estrela, M. J., Olcina-Cantos, J., and Martin-Vide, J.: Future projection of precipitation changes in the Júcar and Segura river basins (Iberian Peninsula) by CMIP5 GCMs local downscaling, Atmosphere, 12, 879, 2021.

Monjo, R., Gaitán, E., Pórtoles, J., Ribalaygua, J., and Torres, L.: Changes in extreme precipitation over Spain using statistical downscaling of CMIP5 projections, International Journal of Climatology, 36, 757–769, 2016.

Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K., and Grover, A.: ClimaX: A foundation model for weather and climate, arXiv preprint arXiv:2301.10343, 2023.

Peral García, M. C., Navascués, B., and Ramos Calzado, P.: Serie de precipitación diaria en rejilla con fines climáticos, 2017.

Pérez, I. A. and García, M. Á.: Climate change in the Iberian Peninsula by weather types and temperature, Atmospheric Research, 284, 106 596, 2023.

Prince, S. J.: Understanding Deep Learning, MIT press, 2023.

Quesada-Chacón, D., Barfus, K., and Bernhofer, C.: Repeatable high-resolution statistical downscaling through deep learning, Geoscientific Model Development, 15, 7353–7370, 2022.

Radford, A.: Improving language understanding by generative pre-training, 2018.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision, in: International conference on machine learning, pp. 8748–8763, PMLR, 2021.

Rampal, N., Gibson, P. B., Sood, A., Stuart, S., Fauchereau, N. C., Brandolino, C., Noll, B., and Meyers, T.: High-resolution downscaling with interpretable deep learning: Rainfall extremes over New Zealand, Weather and Climate Extremes, 38, 100 525, 2022.

Rampal, N., Hobeichi, S., Gibson, P. B., Baño-Medina, J., Abramowitz, G., Beucler, T., González-Abad, J., Chapman, W., Harder, P., and Gutiérrez, J. M.: Enhancing Regional Climate Downscaling through Advances in Machine Learning, Artificial Intelligence for the Earth Systems, 3, 230 066, 2024.

Schmude, J., Roy, S., Trojak, W., Jakubik, J., Civitarese, D. S., Singh, S., Kuehnert, J., Ankur, K., Gupta, A., Phillips, C. E., et al.: Prithvi WxC: Foundation Model for Weather and Climate, arXiv preprint arXiv:2409.13598, 2024.

Sha, Y., Gagne II, D. J., West, G., and Stull, R.: 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, 2020a.

Sha, Y., Gagne II, D. J., West, G., and Stull, R.: Deep-learning-based gridded downscaling of surface meteorological variables in complex terrain. Part II: Daily precipitation, Journal of Applied Meteorology and Climatology, 59, 2075–2092, 2020b.

Soares, P. M., Johannsen, F., Lima, D. C., Lemos, G., Bento, V., and Bushenkova, A.: High resolution downscaling of CMIP6 Earth System and Global Climate Models using deep learning for Iberia, Geoscientific Model Development Discussions, 2023, 1–46, 2023.

Sobolowski, S., Somot, S., Fernandez, J., Evin, G., Maraun, D., Kotlarski, S., Jury, M., Benestad, R. E., Teichmann, C., Christensen, O. B., Katharina, B., Buonomo, E., Katragkou, E., Steger, C., Sorland, S., Nikulin, G., McSweeney, C., Dobler, A., Palmer, T., Wilke, R., Boé, J., Brunner, L., Ribes, A., Qasmi, S., Nabat, P., Sevault, F., Oudar, T., and Brands, S.: EURO-CORDEX CMIP6 GCM Selection & Ensemble Design: Best Practices and Recommendations, https://doi.org/10.5281/zenodo.7673400, https://doi.org/10.5281/zenodo.7673400, 2023.

Spanish Meteorological Agency: Archive of Regionalized Climate Scenarios, Spanish National Adaptation Plan (PNACC), https://archivo-proyecciones-climaticas.aemet.es/, accessed: 2025-07-21, 2021.

Taboada, F., Herrera García, S., Anadón Álvarez, R., Álvarez García, M. A., Colina Vuelta, A., Gutiérrez San Millán, E., Fernández Iglesias, J. C., and Guardado Fernández, C.: Development of High-Resolution Climate Change Scenarios for the Principality of Asturias, Technical report, CuCC-Climate Change Chair, University of Oviedo, https://cucc-uo.es/elaboracion-de-escenarios-de-cambio-climatico-de-alta-resolucion-sobre-el-principado-de-asturias/, collaborators: Consejería de Transición Ecológica (Asturias) and AEMET; Available online, 2024.

Vandal, T., Kodra, E., Ganguly, S., Michaelis, A., Nemani, R., and Ganguly, A. R.: Deepsd: Generating high resolution climate change projections through single image super-resolution, in: Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining, pp. 1663–1672, 2017.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A., and Bottou, L.: Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion., Journal of machine learning research, 11, 2010.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H.: How transferable are features in deep neural networks?, Advances in neural information processing systems, 27, 2014.

Zhu, H. and Zhou, Q.: Advancing Satellite-Derived Precipitation Downscaling in Data-Sparse Area Through Deep Transfer Learning, IEEE Transactions on Geoscience and Remote Sensing, 62, 1–13, 2024.