UNIVERSIDAD DE LA LAGUNA

TRABAJO DE FIN DE GRADO

Climate regionalisation through machine learning techniques

Autor:
Luis Ardévol Mesa

Tutor:
Albano José González Fernández

Facultad de Ciencias
Grado en Física

May 22, 2024

Resumen

Los modelos climáticos globales (MCGs) son herramientas fundamentales en la predicción de los cambios climáticos. Sin embargo, la resolución espacial de los MCGs es limitada, por lo que no son de utilidad en terrenos abruptos o zonas costeras, al no considerar sus características propias. Para solucionar este problema, se utilizan modelos climáticos regionales (MCRs) que permiten obtener información más detallada sobre el clima en una región específica. Sin embargo, los MCRs son computacionalmente costosos.

En este trabajo se propone una metodología para realizar una regionalización climática, utilizando técnicas de aprendizaje automático (método estadístico), que permita la mejora de la regionalización en áreas de orografía compleja. El objetivo será emular la salida de diversas simulaciones dinámicas realizadas por el Grupo de Observación de la Tierra y la Atmósfera de la Universidad de La Laguna con el modelo WRF.

Para ello, se entrenarán dos redes neuronales convolucionales, una para la temperatura y otra para la precipitación, utilizando como predictores datos de reanálisis de ERA-5. Las predicciones de los modelos estadísticos con datos de ERA-5 y de tres modelos globales, para periodos pasados y futuros, se compararán con los resultados de las simulaciones dinámicas de WRF que usan esos mismos datos como entrada.

Los resultados obtenidos muestran que la red neuronal convolucional para la temperatura tiene un buen rendimiento, muy superior al obtenido para las precipitaciones. Además, tomando en cuenta las proyecciones con los modelos globales, se ve reflejada la tendencia esperada siguiendo el escenario RCP 8.5.

Abstract

Global climate models (GCMs) are essential tools for predicting climate change. However, the spatial resolution of GCMs is low, so they are not useful in complex orography or coastal areas, as they do not consider their own features. To overcome this problem, regional climate models (RCMs) are used to obtain more detailed information about the climate in a specific region. However, RCMs are computationally expensive.

In this project we propose a methodology to perform climate regionalisation, using machine learning techniques (statistical method), which allows the improvement of regionalisation in areas of complex orography. The goal will be to emulate the output of several dynamic simulations carried out by the *Grupo de Observación de la Tierra y la Atmósfera* of the University of La Laguna with the WRF model.

For this purpose, two convolutional neural networks will be trained, one for temperature and one for precipitation, using ERA-5 reanalysis data as predictors. The predictions of the statistical models with ERA-5 data and three global models, for past and future periods, will be compared with the results of WRF dynamic simulations using the same data as input.

The results obtained show that the convolutional neural network for temperature performs much better than the one obtained for precipitation. Furthermore, taking into account the projections with the global models, the expected trend following the RCP 8.5 scenario is reflected.

Contents

1 Introduction and study objectives

Resumen

En este capítulo se hará una introducción a los conceptos principales usados en el proyecto. Se plantean los distintos métodos de regionalización, dinámicos y estadísticos. La implementación de redes neuronales en estos métodos es un tema novedoso que presenta buenos resultados en muchos de los contextos en los que ha sido probado. Se profundizará en el concepto de downscaling estadístico, así como sus condiciones ideales de aplicabilidad. Además, se hará una introducción al *deep learning* y un repaso de varios tipos de redes neuronales, centrándonos en las redes convolucionales usadas en este trabajo.

1.1 Basic concepts

This work aims to obtain climate predictions for temperature and precipitation in the Canary Islands using statistical methods. These methods enhance the spatial resolution of Global Climate Models and reduce the computational cost of both Global and Regional Climate Models (GCMs and RCMs).

GCMs are numerical models that simulate the interactions between the main climate components, namely atmosphere, ocean, earth surface and ice, providing predictions for different variables like temperature, precipitation, wind speed, etc. These are computationally expensive models, thus they have a coarse spatial resolution. To overcome this problem, downscaling techniques are used to improve the resolution.

1.1.1 Downscaling. Statistical and dynamic

Climate downscaling refers to a set of methods that improve the spatial resolution given by GCMs for a certain area. Two different techniques are mainly used: statistical and dynamic. The latter develop a higher resolution model using GCMs data as initial and boundary conditions for the region of interest. Analogous to the GCMs, interactions between the main climate components are simulated, preserving their physical principles. As these methods follow the same approach of GCMs, they are computationally expensive. The models thus developed are often called Regional Climate Models. More information about climate downscaling can be found in Ekström et al., 2015.

In contrast, statistical methods aim to establish statistical relationships between a large-scale and coarse resolution dataset of meteorological variables (predictors), and historical records of local observations (predictand), such as temperature and/or precipitation. Specifically, empirical relations linking atmospheric synoptic situations with the

local variables of interest are sought. Once these relationships are established, making predictions on those variables is direct. These methods are less computationally demanding and, despite overlooking the underlying physics of the problem, they have shown a great performance under some assumptions that will be presented below. See Wilby et al., 2014 for more information on statistical downscaling.

On predictors selection, reanalysis data is commonly used, while both high-resolution observation grids and meteorological records on local stations are taken as predictands. Reanalysis data is a blend of observations with past short-range weather forecasts rerun with modern weather forecasting models, and provides the most complete picture currently possible of past weather and climate. Statistical models, as exposed on Baño Medina, 2021, are built on three facts the model's quality will depend:

- Predictors must be informative enough to describe the local variability of the predictand of interest.
- Predictors must be realistically simulated by GCMs. A minimum requirement would be to assure that GCM predictors present no biases with respect to their corresponding reanalysis data for the calibration period, otherwise, the high-resolution products will inherit the biases.
- The model has to show a certain extrapolation capacity with respect to the calibration conditions.

There are three different approaches to predict weather variables (Maraun and Widmann, 2018): Perfect-Prognosis (PP), Model Output Statistic (MOS), and Weather Generators (WG). The last of them pretends to learn the distributional moments of the variable of interest at a particular point from observational records. To make predictions, the parameters are perturbed consistently with the climatic change given by GCM/RCM. MOSs try to find a statistical relation between the predictand simulated by a climate model (either GCM or RCM) with its corresponding observational data. This method is mainly used in bias adjustment techniques[1]. The PP approach is based on learning a predictor-predictand relation using a particular method, for example, generalised linear models (GLMs), taking datasets of observational data for both predictor and predictand.

To assess the performance of the model, a cross-validation scheme is followed, with the most simple approach being the hold-out method. This approach is widely used in the supervised training of any machine learning model, and involves dividing the entire dataset into train and test sets which should be disjoint subsets. The former is used to calibrate the model (i.e., train it), while the latter performs duties as checking the extrapolation capacities of the model. Another approach arises if we add a validation set to the ones already presented. This set is usually taken as approximately 10% of the train set and helps to periodically evaluate the results during the training process. This set allows for easier detection of over and underfitting. Berrar, 2018 gives a brief introduction to the most common types of cross-validation and their related data resampling methods.

[1] Biases can arise due to multiple factors such as limited spatial resolution, simplified processes, and/or incomplete understanding of the global climate system.

One of the most important tasks for the hold-out cross-validation involves how to split the data, because this choice has a major role in the quality of the model's results, as they have to present a low bias and variance. Different ways to do that are tested by Reitermanová, 2010 and Medar et al., 2017. The former is a more general approach, while the latter shows the impact of this data split on time series forecasting.

Many statistical methods, including those based on generalised linear regression and analog techniques, need an overfitting to handle the high dimensionality of predictor data, which can lead to a loss of information. To overcome this problem, deep neural networks have recently been implemented in contrast to classical PP statistical downscaling methods. In this framework, convolutional neural networks (CNNs) have shown an exceptional capacity to manage the high dimensional data for a wide variety of purposes, just as for learning predictor-predictand non-linear relations, as shown on Ashraf et al., 2020 and Li et al., 2023.

1.1.2 Deep Learning. Convolutional Neural Networks

At first, conventional machine learning techniques had a poor ability to process raw data. To overcome this problem, representation learning emerged as a set of methods that enable machines to process this kind of information. Along this line, deep-learning (DL) is a group of techniques or methods of representation learning with multiple representation levels from the simple but non-linear composition of modules. Each of them transforms a certain representation level to a higher level of abstraction. The composition of enough transformations allows the machine to learn quite complex functions. Hereafter, just some of the key concepts for this project will be covered. For a better understanding on deep learning, see Goodfellow et al., 2016.

Before delving into Convolutional Neural Networks (CNNs), let's review the basic structure of a neural network (NN). A basic NN is composed of its fundamental unit, a neuron[2]. Neurons are fed with some input data, each one of them with a certain weight (associated with the input importance) and a bias, used to offset the result and help the model to shift the activation function towards the positive or negative side, thus increasing the level of flexibility in the activation process. To introduce non-linearity to the model and determine if a neuron has a non-zero value, an activation function is needed. Applying the activation function to the weighted sum of inputs plus the bias gives the output of the neuron. Depending on the network's goal, there are different common choices of this particular function, such as hyperbolic tangent or sigmoidal functions for classification, since their values are bounded, or ReLU (Rectified Linear Unit) for linear regression.

Regardless of the activation function used, single neurons cannot achieve a non-linearity degree that allows them to learn complex patterns. Therefore, neurons are bundled into layers. Through this arrangement, complex structures can be achieved, thereby defining the topology of the network; we will further explore this topic later. There are three different types of layers: the input layer takes the input data and passes it to the hidden layers, which

[2]They are called artificial neurons because of its similarities with a real neuron, but they are denoted as neurons for simplicity.

extract and transform the data, learning from it. Finally, the output layer produces the network's output. It can have as many hidden layers as it is required. The NN aims to learn a non-linear relationship between predictor variables, x, and predictand, y, parameterised by a set of coefficients ω that explains the predictand variable,

$$y = f^{\omega}(x) \tag{1.1}$$

The process described earlier, where data flows from the input to the output layer, is called feedforward propagation, and it is the very first step on most network's training. The set of coefficients is updated to fit the data according to a loss function that measures the difference between the output and the expected value for that input data, seeking to minimise the loss. Then, the gradients of the loss with respect to the weights and biases is calculated and propagated backwards to adjust these parameters. This process is called backpropagation, and the method used to adjust parameters is typically a gradient descent method, but it can also be a stochastic gradient descent (SGD) or Adam optimisation, for example. To end up with, the model's outputs must be tested on some data to assess its performance and extrapolation capabilities. It is important to note that the output is strictly related to the chosen loss function.

Considering the network topology, there are multiple types of neural networks. The most common topology is the feedforward neural network, in which neurons are disposed in layers connected sequentially. Convolutional neural networks are a type of feedforward neural network that use convolution kernels to extract higher-level features and avoid vanishing or exploding gradients, seen during backpropagation in some topologies. Convolution allows the network to be deeper by reducing the number of free parameters.

A CNN expects the input as a tensor of shape

$$\textit{number of inputs} \times \textit{height} \times \textit{weights} \times \textit{channels}$$

These networks can typically have three different types of layers: convolutional layer, a pooling layer, and a fully connected layer. The first of them performs a convolution over the input data and passes it to the next layer. This is, essentially, the dot product between a matrix of learnable parameters, also known as kernel, and a matrix representing a restricted portion of the data. The kernel is smaller than the data in width and height, but it is deeper so it extends over the channels. Through movements over the height and width of the image, the kernel is able to produce a two-dimensional (spatial) representation of the features learned by the kernel, called filter map. This mechanism resembles the response of visual cortex neurons to a stimulus.

Pooling layers reduce the dimension of the data by clustering the outputs of multiple neurons in a layer into a single neuron of the next layer. The most common ways of pooling are *max*, where the maximum value of each cluster is taken, and *average*, where the average value is used. Fully connected layers, as their name suggests, connect every neuron from one layer to every neuron in the next layer. That way, it can be computed as in typical neural networks to give the output data.

Because of its ability to extract complex patterns from a big dataset, and the fact that they are able to avoid overfitting, CNNs arise as a particularly suitable approach for statistical downscaling (Baño Medina et al., 2020).

1.2 Study goals

The Canary Islands, as many other regions, have a complex orography to deal with. In addition, in GCMs, due to its coarse spatial resolution, small territories are contained in one or a few grid cells, thus they cannot adequately simulate climatic changes. For the Canary Islands, for example, a unique grid cell can contain two or more islands. In this case, the value shown on this cell will be the mean of all values inside them and, since weather may be completely different on the two islands (even inside the same island), the reflected value will show an accurate prediction for neither island. For this kind of regions, downscaling is a necessary technique to make predictions on climate-related variables.

As exposed earlier, deep learning has been proven to be a useful technique for statistical downscaling. Because of its exceptional way of handling data, convolutional neural networks are a convenient choice for this task.

This project aims to implement two different CNNs to apply a statistical downscaling in a region centered on the Canary Islands. This method is based on the approach exposed in Baño Medina et al., 2020, who applied it to the European continent. The results of temperature and precipitation predictions will be compared with the ones from a dynamic downscaling simulation done by the *Grupo de Observación de la Tierra y la Atmósfera, (GOTA)* of the University of La Laguna for the same region. Due to the extreme difference of computational expense, close results in a certain variable will show that, with the available predictor's data, a reliable prediction can be made from a model overlooking the underlying physics of the problem.

The code used to obtain these results is available at the following GitHub repository `https://github.com/uisitoam/Climate-Regionalisation`, along with all the results presented during this project.

2 Data and methods

Resumen

En este capítulo se introducirán las características de los datos usados en las simulaciones, tanto predictores como predictandos, así como el área que comprenden y una breve descripción de la misma. Además, se especifican las métricas usadas para validar y comparar los resultados obtenidos que se presentarán en el próximo capítulo, junto con el método de validación cruzada usado durante el entrenamiento de las redes. Por último, se exponen los modelos de redes neuronales convolucionales usados para temperatura y precipitación. Se verá que los modelos son muy similares, diferenciándose principalmente en el número de filtros convolucionales, salidas de la red y función de pérdida, esta última seleccionada acorde a la distribución que siguen los datos de temperatura y precipitación.

2.1 Area of study and data

The Canary Islands have a complex terrain which diverges between islands. A common geographic feature is the coastal influence; western islands present mountain-valley systems and are usually more mountainous than eastern islands, thus coexisting numerous microclimates. This orography, along with the reduced size of the territory, hinders specially the precipitation forecasting. Another reason for this is the influence on the predictands by local phenomena which are not captured by the predictors.

The VALUE COST Action (2012-2015) is an European project developed by several research centers with the aim of setting "standards" in downscaling methods to ease the comparison of results between the different groups (Maraun et al., 2015). Following the suggestions of this project, 20 standard predictors from the ERA-5 reanalysis are used. [1] In particular, the data used as a predictor will have four dimensions, namely time, level, latitude and longitude. Latitude data is comprehend between 22 and 40 degrees-North, increasing by two degrees each time, and so does longitude, although this one spans from -32 to 0 degrees-East. This way, a 10×17 grid is used as spatial data, while the covered area is shown in Figure I. Additionally, the level variable adds a “height” dimension; the spatial grid mentioned before has values at three different air pressure levels: 500, 700 and 850 mbar.

The variables used as predictors are the geopotential z, given in $[m^2/s^2]$, specific humidity q in water kilograms per air kilograms, $[kg/kg]$, the air temperature t in Kelvin

[1] ERA-5 data was downloaded from the Copernicus Climate Change Service (2022), DOI: 10.24381/cds.e2161bac

and the u and v components of the wind, measured in $[m/s]$. All this data is available each day from the very start of 1982 to the end of 2019. Thus, the predictor available will have dimension

$$\textit{(time, level, latitude, longitude)} = (13392, 15, 10, 17)$$

FIGURE I: Area covered by the predictors data.

FIGURE II: Area covered by the predictands data.

The predictands here considered are daily mean temperature and accumulated precipitation. The data used correspond to a climate regionalisation simulation performed by the WRF (Hersbach et al., 2020) mesoscale model (Skamarock et al., 2019) using ERA-5 reanalysis data as initial and boundary conditions. These regionalised data were obtained by the *Grupo de Observación de la Tierra y la Atmósfera* (GOTA) at ULL, using the WRF configuration defined in previous works (Pérez et al., 2014; Expósito et al., 2015). This data defines a grid over the Canary Islands, ranging from 27.5 to 29.3 degrees-North in latitude and from -18.2 to -13.4 degrees-East in longitude. This coordinates gives a grid of dimension *(latitude, longitude) = (68, 158)*, and the covered area is shown in Figure II. Although different configurations were used throughout the work, finally only values

corresponding to land grid points were considered as predictands. The available data have a value for this grid each day since 5th of January of 1982 until the end of 2019. Altogether, the predicted variables will have a time dimension of 13392.

Once the networks have been trained using data corresponding to the recent past (from 1982 to 2006 as train set and from 2006 to 2009 as validation set), the ability of the obtained systems to emulate the behaviour of other climate simulations is analysed. Thus, the results of other climate regionalisations carried out by GOTA and driven by three GCMs belonging to CMIP5 (Coupled Model Intercomparison Project, phase 5) were considered. Specifically, results of r1i1p1 simulations of GFDL-ESM2M, IPSL-CM5A-MR and MIROC-ESM were used as initial and boundary conditions for WRF simulations (Perez et al., 2022). They correspond to three 30-year periods (1980-2009, 2030-2059, 2070-2099) and were performed with the results of the GCMs using the Representative Concentration Pathway (RPC) 8.5 (Taylor et al., 2012).

2.1.1 Evaluation indices and cross-validation

Metrics	Variable	Units
Bias (for the mean)	temp., precip.	°C, %
Bias (for the 2nd percentile, P2)	temp.	°C
Bias (for the 98th percentile, P98)	temp., precip.	°C, %
Root mean square error (RMSE)	temp., precip.	°C, mm/day
Ratio of standard deviations	temp.	ad
Pearson correlation	temp.	ad
Spearman correlation	precip.	ad
Bias (warm annual max spell, WAMS)	temp.	day
Bias (cold annual max spell, CAMS)	temp.	day
Bias (wet annual max spell, WetAMS)	precip.	day
Bias (dry annual max spell, DryAMS)	precip.	day

TABLE 2.1: VALUE metrics used to validate downscaling methods considered (see Table 2). "ad" denotes adimensionality.

In this work, a subset of VALUE metrics is considered to assess the performance of the implemented models, as it is done by Baño Medina et al., 2020 (summarised in Table 2.1).

All in all, biases in seven different metrics are computed: five for temperature and four for precipitation. The bias in the mean, the most common of them, is evaluated for both variables, together with the bias in the 98th percentile (P98). Additionally, the bias in the 2nd percentile (P2) is used for temperature. These three biases are given as absolute differences for temperature, in °C, and as relative differences for precipitation (in %). Biases for four temporal indices are also calculated, two of them for each predictand. For temperature, the median warm (WAMS) and cold (CAMS) annual maximum spells are used, while for precipitation the median wet (WetAMS) and dry (DryAMS) annual

maximum spells are useful for this task.

A convenient measure to look at is the differences between predicted and predictand values, consequently, the root mean square error (RMSE) is evaluated for both temperature and precipitation. For the latter, just the observed wet days (rainfall $>$ 1 mm) are used. The ratio of standard deviations is computed just for temperature. Correlation is also a metric to be considered; two different correlation coefficients have been calculated: the Pearson coefficient for temperature and the Spearman rank one for precipitation, due to its non-Gaussian nature.

A cross-validation scheme must be followed to assess the performance of the models. A hold-out method is used for both temperature and precipitation models. As pointed in subsection 1.1.1, there are different methods to split the data for a hold-out cross-validation. Convenience sampling is the approach taken in this work, due to its suitability and efficiency when dealing with time series. Systematic sampling could be another approach, however, it is hard to find an appropriate ordering for most datasets. In the method here considered, the dataset is split into discrete blocks of time intervals. Specifically, the data is divided in three subsets: the period between 1982 and 2006 is considered as the train set, the data from 2007 to 2010 is taken as the validation set, and the test set is composed of the period between 2011 to 2019. This sampling is chosen on account of the warmer conditions shown in these last years: a critical feature of the models developed is their extrapolation capabilities, thus a time spell with slightly different conditions is selected as a test set.

2.2 Methodology

As it was advanced in section 1.2, two different CNNs will be implemented to develop a statistical downscaling method for each of the predictands. As the predictors used come from the GOTA simulations using the WRF mesoscale model, the main purpose will be to emulate these results. Both CNNs are similar, but the layers, along with the output values, have differences. This section aims to shed light into the structure and behaviour of these neural networks. For the CNN models used in this work, the best-performing topology developed in Baño Medina et al., 2020 is implemented for precipitation, while for temperature the considered model is exposed in Baño Medina et al., 2022.

Neural networks are a stochastic machine learning algorithm, as the initial weights are random, thus the model will learn from a different starting point each time. Due to this fact, training a model will lead to different results each time, even with the same train and test datasets. In order to account this variability and obtain solid results, the networks presented afterwards will be trained with the predictors five times, and the results given in the next chapter will show the median results. The appendix B shows the numerical results for the median and the mean along with the standard deviation of the five runs.

2.2.1 Temperature model

Layer (type)	Output Shape	Param #	Connected to
input_layer (InputLayer)	(None, 10, 17, 15)	0	-
conv2d (Conv2D)	(None, 8, 15, 50)	6800	input_layer[0][0]
conv2d_1 (Conv2D)	(None, 6, 13, 25)	11275	conv2d[0][0]
conv2d_2 (Conv2D)	(None, 4, 11, 10)	2260	conv2d_1[0][0]
flatten (Flatten)	(None, 440)	0	conv2d_2[0][0]
dense (Dense)	(None, 1059)	467019	flatten[0][0]
dense_1 (Dense)	(None, 1059)	467019	flatten[0][0]
concatenate (Concatenate)	(None, 2118)	0	dense[0][0], dense_1[0][0]
Total params: 954373			
Trainable params: 954373			
Non-trainable params: 0			

TABLE 2.2: Detailed architecture of the Convolutional Autoencoder of the temperature model.

This model is mainly composed of three convolutional layers (with 50, 25 and 10 filters/kernels respectively) and two concatenated dense layers which behave as the output layer. The former layers use 3D kernels of shape 3×3 to convolute the raw data provided by the input layer. These layers have a 'valid' padding, thus the output feature map of each one of them is smaller than the input feature map and its spatial resolution is reduced.

A flatten layer to reshape the data-driven spatial feature in order to feed the dense layers is also required. Dense layers have a number of neurons equal to the dimension of the predicted data; in short, each land grid box uses two output neurons, as the output layer comprises two dense layers. The explicit model is shown in Table 2.2.

Loss function presents a major role in the output values. Training aims to optimise the negative log-likelihood of a Gaussian distribution, since the predictand is assumed to follow this distribution. Therefore, the output of the network corresponds to the mean and variance of the distribution, i.e., the distributional parameters.

Any loss consisting of a negative log-likelihood is a cross-entropy between the empirical distribution defined by the training set and the probability distribution defined by model. It is proved in Goodfellow et al., 2016 and shown in Appendix A that the approach here chosen is equivalent to use the mean square error (MSE) as loss function, since MSE is the cross-entropy between the empirical distribution and a Gaussian model.

2.2.2 Precipitation model

Layer (type)	Output Shape	Param #	Connected to
input_layer (InputLayer)	(None, 10, 17, 15)	0	-
conv2d (Conv2D)	(None, 10, 17, 50)	6800	input_layer[0][0]
conv2d_1 (Conv2D)	(None, 10, 17, 25)	11275	conv2d[0][0]
conv2d_2 (Conv2D)	(None, 10, 17, 1)	226	conv2d_1[0][0]
flatten (Flatten)	(None, 170)	0	conv2d_2[0][0]
dense (Dense)	(None, 1059)	181089	flatten[0][0]
dense_1 (Dense)	(None, 1059)	181089	flatten[0][0]
dense_2 (Dense)	(None, 1059)	181089	flatten[0][0]
concatenate (Concatenate)	(None, 3177)	0	dense[0][0], dense_1[0][0], dense_2[0][0]
Total params: 561568			
Trainable params: 561568			
Non-trainable params: 0			

TABLE 2.3: Detailed architecture of the Convolutional Autoencoder of the precipitation model.

This model is composed of three convolutional layers (with 50, 25 and 1 filters/kernels respectively) and three concatenated dense layers which behave as the output layer. The former layers use 3D kernels of shape 3×3 to convolute the raw data provided by the input layer. A flatten layer to reshape the data-driven spatial feature in order to feed the dense layers is also required. Each land grid box uses three output neurons, as the output layer comprises three dense layers. The explicit structure of this model is shown in Table 2.3.

Thus far, the only differences between temperature and precipitation model are the number of filters of the third convolutional layer (10 for temperature vs 1 for precipitation) and the number of dense layers building the output layer. Furthermore, these convolutional layers have a 'same' padding, so there is no reduction in the spatial resolution of the feature maps.

An appropriate probability density function must be selected to represent the distribution of precipitation. Williams, 1997 suggested using a mixed Bernoulli-Gamma distribution for describing precipitation series that include both days with no precipitation and days with precipitation. This leads to an additional distinction: the loss function. Training this model optimises the negative log-likelihood of a Bernouilli-Gamma distribution, following the approach previously introduced by Cannon, 2008, who showed that Bernouilli-Gamma and Poisson-Gamma distributions can fit precipitation series that include both dry and wet days. As a result, the output of the network corresponds to the probability of rain, shape and scale factors (p, α, β).

Let's delve into how the precipitation amount is obtained with this model. The Bernouilli-Gamma probability density function is given by

$$f(y;p,\alpha,\beta) = \begin{cases} 1-p & \text{for } y = 0 \\ \dfrac{py^{\alpha-1}e^{-y/\beta}}{\beta^{\alpha}\Gamma(\alpha)} & \text{for } y > 0 \end{cases} \tag{2.1}$$

where $0 \leq p \leq 1$, $\alpha > 0$ and $\beta > 0$ are the distributional parameters mentioned before and y is the precipitation amount. The mean value of the distribution is $\mu = \alpha \cdot \beta$.

The amount of rainfall in a certain day, i, is the mean of its distribution, $r_i = \alpha_i \cdot \beta_i$, and then by probability of occurrence, values where there is no rainfall at all (under a given value) are cut down to zero. To fix this threshold for the rain occurrence one must calculate the non-rainy days percentage from the predictand data. Out of the training outputs of the network, the value of probability p for which the same number of non-rainy days are obtained, i.e., a quantile, is the desired threshold.

While this method gives a deterministic value, it can also provide stochastic predictions by simulating a random value from the distribution.

3 Results

Resumen

En este capítulo se presentan los resultados obtenidos con las redes convolucionales expuestas para temperatura y precipitación. Se analiza cada uno por separado, realizando una discusión detallada de los resultados arrojados haciendo uso de los datos de los tres modelos climáticos globales, así como de los datos de reanálisis de ERA-5. Para esto, se usan las métricas de VALUE expuestas en la tabla 2.1. Además, se comparan las predicciones de temperatura media y cantidad de precipitación con la de WRF para todo el conjunto de datos (MCGs y ERA-5) considerado. Esta comparación se acompaña de una discusión de la validez de los resultados en el contexto del escenario RCP 8.5.

3.1 Temperature model

From a general point of view, the temperature prediction model emulates the results given by the WRF model in a similar way in all the time ranges considered, together with the data from the three global models and ERA-5. The further out the time period considered, however, the more deviant the projections become.

Figure III contains the predictions used to evaluate the model's performance along with the projections using GCMs. As the goal is to emulate the outputs of WRF, the CNN results using a given model (either ERA-5 or GCMs) are compared with WRF predictions using the same data of the same model. For the sake of brevity and conciseness, hereafter, the predictions made with CNN or WRF model, using GCM (or ERA-5) data, will be referred to by the name of the corresponding GCM.

The boxplots displayed in Figure III show the evaluated metrics for all prognoses using the CNN-temperature model. First, the given metric for each land point is obtained. Then, the median and statistics presented on the boxplot consider all land points over the grid.

3.1.1 Performance

In order to evaluate the performance of the model, reanalysis data from ERA-5 is used. As it was stated in the methodology, the models are trained with ERA-5 data from 1982 to 2009 (train and validation sets). The performance of the model will be tested with ERA-5 data from 2010 to 2019. These results will be contrasted with the ones shown on Baño Medina et al., 2020 for Europe (from now on, CNN10 model) and on Guinea et al., 2023 for the Canary Islands (from now on, AEMET model).

Generally speaking, a salient feature of the predictions from ERA-5 reanalysis data is the low dispersion shown by the values between the 25th and 75th percentiles in all considered metrics, as well as the whiskers' extension, which is also lower than that of the other models presented on the same plot.

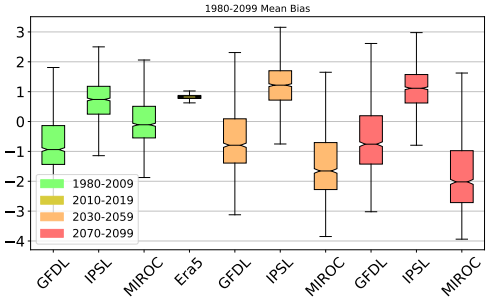
The bias with respect to the mean stands at less than $1°C$, while the bias with respect to the 2nd percentile is slightly higher than 1 degree and the one with respect to the 98th percentile is just under $0°C$ (Figure IIIA, IIIB and IIIC, respectively). This means that the predictions are overestimating the mean and the lower temperatures, while underestimating the higher ones. This behaviour is the same as temperature results of CNN10, but the values presented there are lower than the CNN results for these metrics. However, the dispersion of the data is approximately the same. AEMET model shows an underestimation of both extremes, worse than the results displayed on the analysed figures.

The prognoses with ERA-5 are able to maintain a linear correlation between the results with an R of more than 0.9 (Figure IIID), while clearly reducing the scatter of the data with respect to those predicted by WRF (Figure IIIE). Both metrics are higher in CNN10, showing a similar data dispersion to the one with the CNN model. AEMET presents a similar value for the Pearson correlation coefficient to the one shown by the CNN, but the standard deviation ratio is not calculated. RMSE is akin in all models under review: CNN10 was able to reach values under one degree, whereas the CNN model and AEMET stands at $1.5°C$ (Figure IIIF).

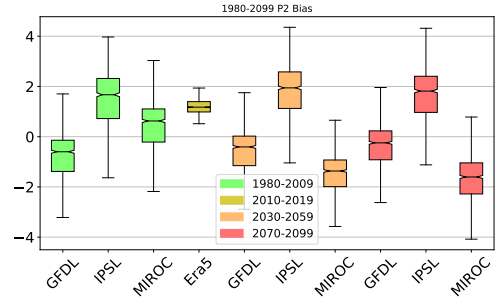
The predictions of cold and warm spells accomplished by CNN10 are impressive, with a median bias of zero days and a scatter of ± 1 day. The CNN model performs similarly in terms of scatter, but the median here is found to be 1.4 days for both of them (Figure IIIG and IIIH, respectively). AEMET does not evaluate these metrics.

Overall, CNN predictions with ERA-5 are able to emulate WRF predictions adequately with the same data. The results obtained with CNN are worse than those shown by CNN10, although it must be mentioned that the latter gives an output grid with a resolution of 0.5 degrees (pixels of approximately 55×55 km), while CNN gives a grid with pixels of approximately 3×3 km. On top of that, the data used by CNN10 is averaged, which makes the values much smoother, simplifying the work of the network.

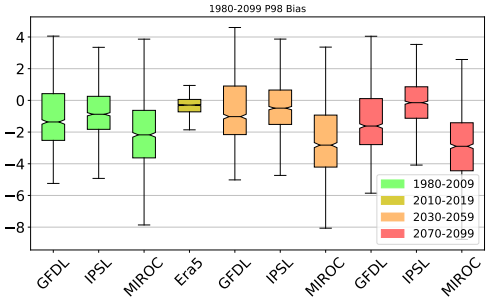
Comparisons with point observations in the Canary Islands can be done with the results of AEMET. These, however, are not arranged in a grid and, moreover, are obtained from different methods, including non-convolutional neural networks. The results obtained with the CNN model outperforms those of AEMET in terms of biases with respect to the extremes, 2nd and 98th percentile. Despite this, some values are found to be similar, and the dispersion of the data is not shown better in any of them.



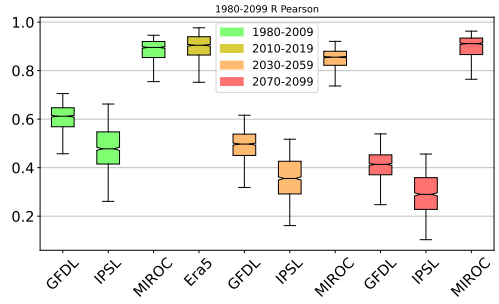
(A) Bias with respect to the mean (°C).



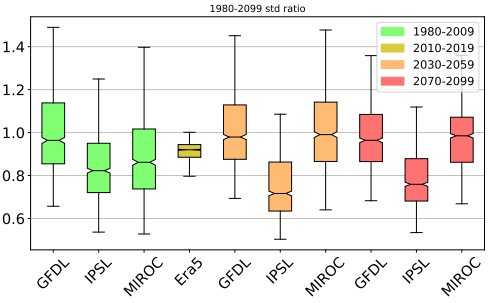
(B) Bias with respect to the 2nd percentile (°C).



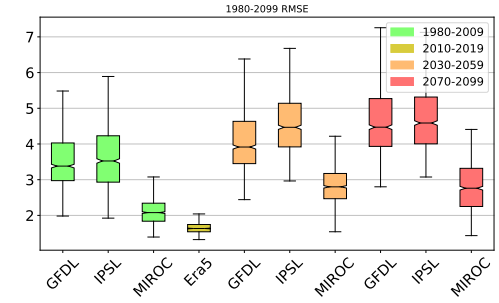
(C) Bias with respect to the 98th percentile (°C).



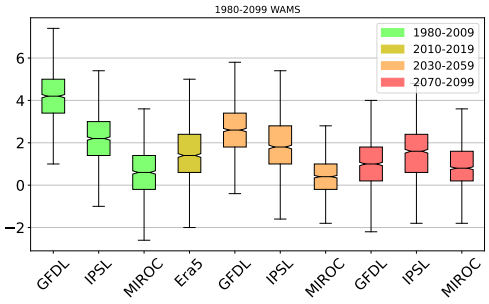
(D) Pearson correlation coefficient.



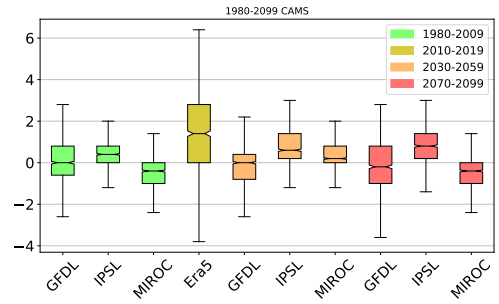
(E) Standard deviation ratio.



(F) Root mean square error (°C).



(G) Bias on WAMS (days).



(H) Bias on CAMS (days).

FIGURE III: Metrics comparison between each model used to make predictions on temperature with the CNN 2.2 with respect to WRF. Four time periods are considered: 1980-2009, 2010-2019, 2030-2059 and 2070-2099. The time period between 2010-2019 is only predicted with ERA-5 data, while the others are obtained from the GCMs models. Values inside the boxes comprise the 25th ($Q1$) to 75th percentile ($Q3$). Inside each box, the median is shown as a line, while the notches represent the confidence interval around the median. The lower whisker extends to $Q1 - 1.5 \cdot IQR$ and the upper whiskers to $Q3 + 1.5 \cdot IQR$, where $IQR = Q3 - Q1$. The points outside these whiskers are considered outliers.

3.1.2 Projections

The bias with respect to the mean is displayed in Figure IIIA. In the period from 1980 to 2009, the best value is the one obtained from predictions made with MIROC data. Nevertheless, the projections with these data deviate significantly from the results of WRF in future periods. For these time spans, the prognoses with the GFDL model, GFDL, become more accurate. The use of these two models leads to an underestimation of this magnitude: the deviation with the MIROC model is just over 0 degrees in the earlier period, and rises in future periods up to almost two degrees, whereas with the GFDL model it remains slightly under one degree in the three time intervals. Although the predictions with the IPSL model are not the most accurate ones, they maintain similar values (approximately 1 degree of overestimation) in the three periods considered.

An analogous behaviour is observed for the bias with respect to the 2nd percentile (Figure IIIB): the predictions with the data from the GFDL and IPSL models are fairly constant throughout the three time spans, while the projections with the MIROC model tend to deviate over time. Here, the GFDL model is the best in all time periods considered, underestimating the WRF results within one degree, whilst the one giving the furthest values from those predicted by WRF is the IPSL model, drifting almost two degrees above. Under the MIROC model, yields are more variable: it overestimates the lowest outputs by about one degree in the foregone period, whereas over future periods it underestimates them by more than one degree.

Notwithstanding the poor emulation of the results for the lowest temperatures, the IPSL model outperforms the two others in terms of predicting the highest values (Figure IIIC). Regarding the bias with respect to the 98th percentile, the projections with the IPSL model are the closest to those given by WRF in the three time periods considered, while the MIROC model is the worst performer. In this case, the use of all the models leads to an underestimation of the predictions given by WRF in the three time spans.

In spite of this comparison, the projections with the three models for the three time periods considered are reasonably close to expectations, both for the mean and the extremes. For all cases, the maximum deviation is found to be of $2^{\circ}C$. The IPSL model overestimates both the mean and the lower end, while the GFDL and MIROC models underestimate these magnitudes (excluding MIROC in the period from 1980 to 2009). All models underestimate the upper end, ranging from 0.1 to almost three degrees Celsius. This entails a problem for the prediction of extreme events and hazardous to the population, such as heat waves.

Pearson correlation is shown in Figure IIID. The MIROC model is the one that best correlates between the projections, hovering around a linear correlation coefficient (Pearson) of 0.9 in the three time periods considered. This metric does not decrease with time, as happens with biases of the MIROC model. The correlation with the other two models is gradually lost over time reaching values of 0.3 in the interval from 2070 to 2099. It is worth noting that, for GFDL and IPSL, the correlation is relatively low from the first period to the last.

Something that should be expected when predicting with the here developed convolutional models is a reduction (or at least a maintenance) of the dispersion of the results, so that the ratio between the dispersion of these predictions with respect to those given by WRF is less than 1 (Figure IIIE). This is fulfilled by using all models in all time periods considered. Under the IPSL model, the networks are able to reduce the dispersion to a greater extent than with the other two models. This reduction is smaller in future time spans with the MIROC model, while in the period from 1980 to 2009, the GFDL model is the worst performer.

The root mean square error (RMSE), shown in Figure IIIF, turns out to be lower with the projections using the MIROC model data, remaining between two and three degrees in all three time intervals considered. The RMSE becomes higher with the other two models, ranging between three and five degrees. While both present a significant difference with the results relying on the MIROC model, those using the IPSL model ranked the worst over the three periods.

The use of the MIROC model is the most suitable over the whole range from 1980 to 2099 as a means to predict the annual maximum warm spell. Projections with this model overestimate the duration of these events by less than a day. This is clearly seen in Figure IIIG. The other two models also yields to an overestimation of the length of these warm spells: whilst the IPSL model predicts a duration of almost two more days, the prognoses with the GFDL model depend on the time period considered; from 1980 to 2009, there is an underestimation of just over four days, in the near future this deviation is greater than two days, and in the far future the overestimation is reduced to just over one day.

Conversely, the prediction of annual maximum cold spells is more accurate than that of the warm spells, maintaining all biases below a single day (both overestimation and underestimation), as displayed in Figure IIIH. For this metric, the use of the GFDL model becomes the best option, leaving the MIROC model behind. The former is only biased in the more distant future, from 2070 to 2099, underestimating the WRF prediction by less than 0.2 days, while the MIROC model has a bias of just under half a day in each of the periods, overestimating the WRF projections in the near future (2030-2059) and underestimating them in the other two. The worst predictions are obtained with the IPSL model, which underestimates in all the periods considered, ranging from just under half a day in the past to almost a day in the far future.

It should be noted that the cold period prognosis is in agreement with those of the bias with respect to the second percentile, where GFDL model usage outperforms the other two models. In stark contrast to this, it is remarkable that this behaviour does not prevail in the results of the warm periods and the bias with respect to the 98th percentile. In the latter, the use of the MIROC model was the worst option due to an underestimation of between 2 and 3 degrees in all time intervals, while with the IPSL model, the predictions emulated the WRF projections better, as the bias approached zero over time. In contrast the use of the MIROC model allows for better predictions of warm spells, as it overestimates the duration of these events by less than a day, as opposed to the IPSL model, which leads to an overestimation of approximately two days in all periods considered.

ERA-5 results can be qualitative compared with GCMs results. The biases with respect to the mean and the 98th percentile are less than $1°C$, the former resulting in an overestimation whilst the latter in an underestimation. At the lower end (2nd percentile), this deviation results in an overestimation of just over one degree. The predictions with this data are able to maintain a linear correlation between the results with an R of more than 0.9, while clearly reducing the scatter of the data with respect to those predicted by WRF. RMSE is found to be less than two degrees. The projections of cold and warm spells are consistent and present the same value for both, being the worst for cold spells.

Based on the values of the eight metrics considered, the use of the MIROC model is the one that gives the best results in predicting the past period, with the GFDL model showing shortcomings with respect to the other two. On the contrary, in future periods, the use of the latter becomes the best option, improving its performance the further ahead it is predicted. The MIROC model significantly worsens its performance in these periods, while the IPSL model maintains similar values in the three time spans considered. Nonetheless, the use of the MIROC model in the interval from 1980 to 2009 results in very close predictions to those sought. The values obtained with ERA-5 reanalysis data are fairly good and reflect the good performance of the model.

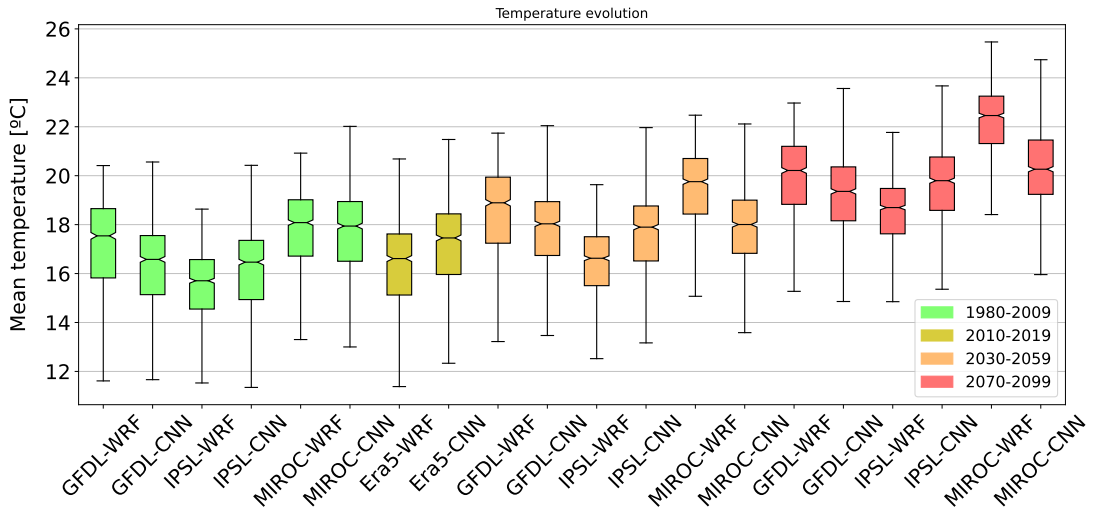


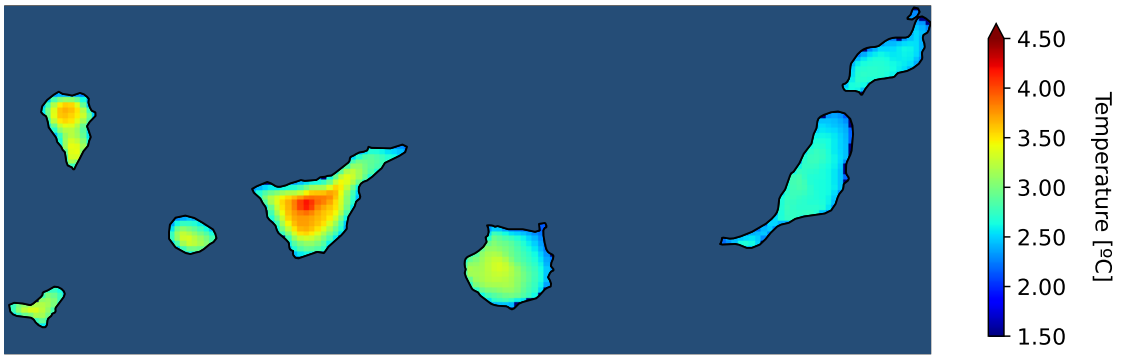
FIGURE IV: Predictions on mean temperature for each model considered in this work. Results for CNN and WRF models are shown for four time periods: 1980-2009, 2010-2019, 2030-2059 and 2070-2099. The time period between 2010-2019 is only predicted with ERA-5 data, while the others are obtained from the GCMs models considered.

Global climate models data in the RCP 8.5 scenario is being used. This scenario envisages an increase in greenhouse gas emissions throughout the 21st century. Under these conditions, an increase of approximately 3.7 degrees is expected by the end of the century. Therefore, it would be desirable that this trend would be mirrored in the predictions made by the statistical model developed in this work with the data from the global models (Figure IV). See Riahi et al., 2011 for more information on this scenario.

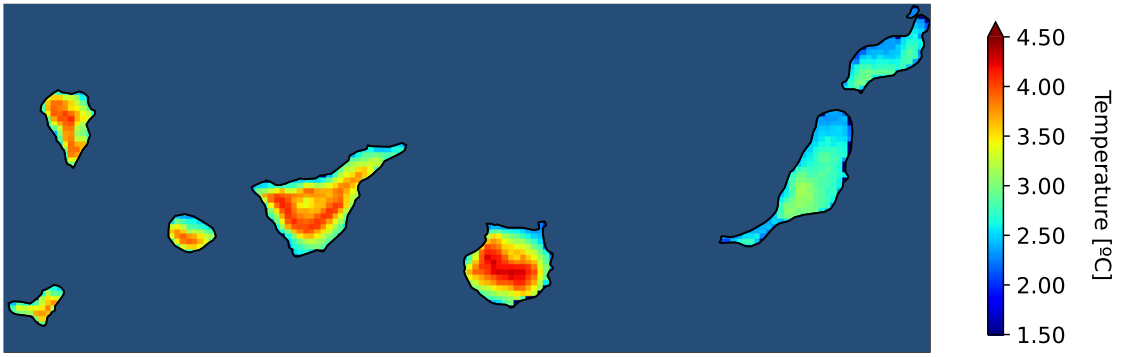
An average temperature between 16 and 18 degrees is observed (depending on the model considered for prediction) in the period from 1980 to 2009, which, considering the bias of the WRF model with respect to the observations, and the bias of the CNN model with respect to WRF, is consistent with the historical data (Perez et al., 2022). Furthermore, the predictions given by CNN are comparable to those given by the WRF model, both in terms of mean and variability. However, the values of the CNN are slightly higher than those of WRF when the IPSL model is used (the same is true for ERA-5), and slightly lower when the other two models are used. This trend is preserved over the three periods considered.

An important point to note is the difference between the results of WRF and CNN: for the GFDL and IPSL models, this difference remains almost constant over time, while with the MIROC model the difference between both predictions increases over time. This behaviour is consistent with that observed for the biases of these three models.

Also noteworthy is the increase in the average temperature over time depicted in Figure IV. While this was to be expected, as discussed above, the temperature increase assumed by the RCP 8.5 scenario is clearly reflected in the predictions, both from WRF and CNN (the one we are most interested in). In the period from 2030 to 2059, the projections of each model are almost 2 degrees higher than in the past period, whilst the increase in the period from 2070 to 2099 is almost 4 degrees with respect to the period from 1980 to 2009.



(A) Predictions of WRF with GFDL data.



(B) Predictions of CNN with GFDL data.

FIGURE V: Temperature dependence on elevation. The difference between the period 2070 to 2099 and 1980 to 2009 is shown.

Future projections for Canary Islands, based on dynamic regionalisations, predict a greater temperature increase at higher elevations (Expósito et al., 2015, Perez et al., 2022). This is in agreement with the so-called elevation-dependent warming, which is typical of tropical and subtropical regions, and consistent with wet adiabatic stratification in these areas, resulting in a reduced lapse rate in the future (Bony et al., 2006). This has mainly biological and ecological implications, but it is of interest to find out whether the developed model is able to replicate this behaviour.

Figure V shows the difference between the results for the period 2070 to 2099 minus 1980 to 2009. In this case, the GFDL model is used, although the use of the three GCMs shows the sought-after behaviour. It can be seen that the distributions predicted by WRF (Figure VA) and CNN (Figure VB) are similar, both depicting a greater increase in temperature with altitude. However, this increase is more sharp in the CNN results. The most remarkable facts are the difference at Teide, where values closer to those given by WRF would be expected, and the large difference in the central area of Gran Canaria, which show higher values than the expected ones. Despite the differences between the two models, it appears that, overall, the CNN model is able to replicate the WRF outcome.

3.2 Precipitation model

Just as in the case of the temperature results, the precipitation prediction model emulates the results given by the WRF model in a similar way in all the time spans, although they deviate more the further the time period considered. However, in this case, the results are not emulated adequately, and variability is lost. In Figure VI all these results are shown. For precipitation, the bias with respect to the mean (Figure VIA) and with respect to the 98th percentile (Figure VIB) is calculated as a percentage, relative to WRF results,

$$\text{Bias} = \frac{\text{CNN prediction} - \text{WRF prediction}}{\text{WRF prediction}} \cdot 100 \tag{3.1}$$

A drawback arising from this definition is that, due to the low precipitation amount in the Canary Islands, the 98th percentile of the predictions turns out to be close to zero on several occasions, so the bias with respect to this percentile is extremely large. This is clearly reflected in Figure VIB, so this metric will not be considered in GCMs results.

3.2.1 Performance

Figure VI contains the predictions used to evaluate the model's performance along with the projections using GCMs. For this predictand, the same scheme as for temperature is followed, and the same statements stands for the boxplots displayed in Figure VI.

ERA-5 results allow us to evaluate the performance of this model. The low dispersion shown by the values between the 25th and 75th percentiles in temperature is not as salient for precipitation. Both biases with respect to the mean and the 98nd percentile encompass a broad set of values.

The bias with respect to the mean stands at 50%, while the bias with respect to the 98th percentile is slightly higher than 70% (Figure VIA and VIB, respectively). This means that the predictions are overestimating the mean and the heavier precipitations. Despite that high values, the underlying cause of them comes from an overestimation of precipitation on the driest regions, as a little deviation in absolute value will suppose a high one in percentage. This behaviour is the contrary to results on Baño Medina et al., 2020, which underestimate the predictions in a very low percentage. Guinea et al., 2023 do not show the extreme, but show a similar bias with respect to the mean as the one given by the CNN.

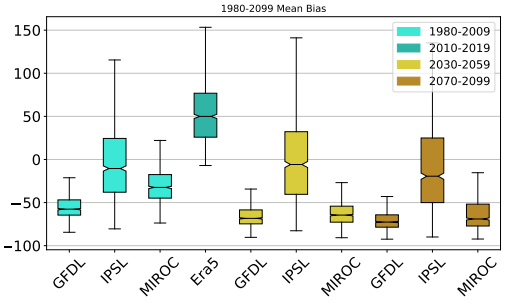
The prognoses with ERA-5 are able to maintain a certain correlation between the results with an Spearman coefficient of more than 0.45 (Figure VIC). This metric is higher in Baño Medina et al., 2020, showing a similar data distribution to the CNN model. A virtually zero Spearman correlation is shown in Guinea et al., 2023, which means the models there developed do not have the ability to maintain the correlation between results. RMSE in wet days is akin in Baño Medina et al., 2020 and CNN models: the former reaches values of 5 mm/day whilst the latter is just under 3 mm/day (Figure VID).

The predictions of wet and dry spells accomplished by Baño Medina et al., 2020 are quite accurate, with a median bias of 0.5 days and a scatter of ± 1 day for the former, and a bias of 2 days with a scatter of ± 2 days for the latter. While predicting wet spells, the CNN model performs better in terms of scatter, but the median here is found to be 3.5 days. With this model, dry spells are underestimated by approximately 10 days, which makes it the worst performer (Figure VIE and VIF, respectively). Guinea et al., 2023 do not evaluate these metrics.

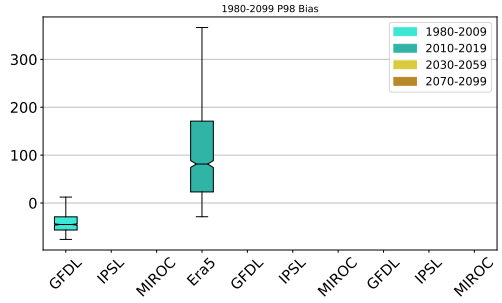
In some ways, the worse performance of the CNN compared to the Baño Medina et al., 2020 can be explained in the same way as in the case of temperature. A further reason could be the large variability of this predictand in the area under study. Guinea et al., 2023 do not evaluate most of the metrics used in this work, but they show a behaviour similar to the CNN model in terms of biases with respect to the mean, while showing not correlation at all between results, unlike the CNN ones.

3.2.2 Projections

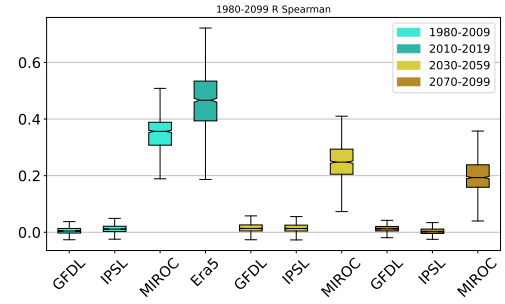
The bias with respect to the mean (Figure VIA) obtained with the IPSL model is by far the best in the three periods. Nevertheless, the standard deviation considering the 5 trainings of this model is the highest, standing at 186%, which detracts from the validity of this value (exact values are shown in Appendix B, tables B.6, B.7 and B.8). The GFDL model leads to consistent results in the three periods (there is a slight increase over time), underestimating precipitation by 50% to 70%. The MIROC model, on the other hand, underestimates precipitation by 30% in the interval from 1980 to 2009, but in the future it shows a similar behaviour to that exhibited with temperature, diverging more from the results of WRF the further the period considered. However, now the difference in values is not as pronounced.



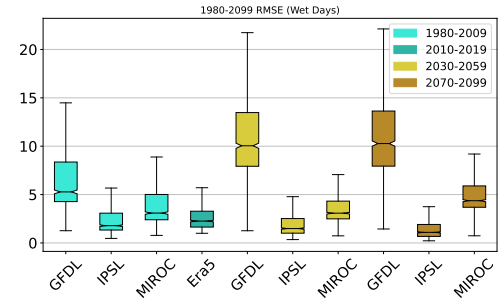
(A) Bias with respect to the mean (%).



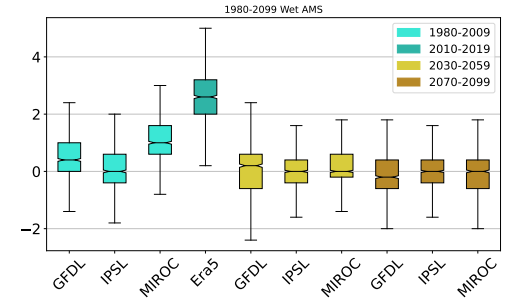
(B) Bias with respect to the 98th percentile (%).



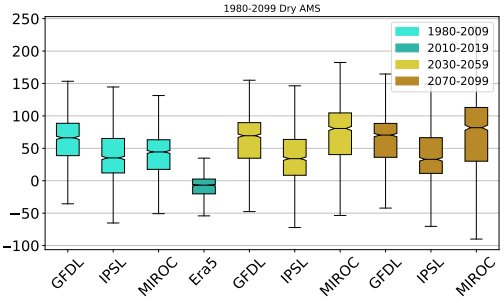
(C) Spearman correlation coefficient.



(D) RMSE only for wet days (mm/day).



(E) Bias on Wet AMS (days).



(F) Bias on dry AMS (days).

FIGURE VI: Metrics comparison between each model used to make predictions on precipitation with the CNN 2.2. Four time periods are considered: 1980-2009, 2010-2019, 2030-2059 and 2070-2099. The time period between 2010-2019 is only predicted with ERA-5 data, while the others are obtained from the GCM models. Predictions on precipitation are so low that, usually the 98th percentile is zero, so the relative P98 Bias is undefined. Boxplots' range of values is defined as in the case of temperature.

Predictions with global models underestimate WRF results to a greater or lesser extent, whereas if ERA-5 data is used, predictions are overestimated by 50%.

Spearman correlation coefficient (Figure VIC) has virtually zero values when using the GFDL and IPSL models. As this coefficient measures how well a relationship between two variables can be described using a monotonic function, this fact implies non-correlation (but not independence) between the variables in the three periods considered. Conversely, this coefficient ranges between 0.35 and 0.2 over time when using the MIROC model, from the highest value for the past period to the lowest for the farthest future. In this case, a slight positive correlation or association between the variables is demonstrated.

The root mean square error shown in Figure VID, considering only days with precipitation greater than 1 mm/day (that is what is called a wet day), displays very low values in the predictions obtained with the IPSL and MIROC models. The former is the best, with values around 1.5 mm/day in the three periods, while the latter is a step behind with values between 3 and 4 mm/day. In the interval from 2070 to 2099, the results with MIROC increase slightly compared to the other two periods with this model. The results using the GFDL model are, by far, the worst: from 1980 to 2009, it presents a value slightly above 5 mm/day, whereas in future time spans this value soars to approximately 10 mm/day. This indicates the magnitude of the error in the prediction with each model. The results using ERA-5 data present an RMSE as low as those with the IPSL model.

Wet annual maximum spells (Figure VIE) are predicted almost perfectly with all three models in the three periods considered, with the greatest deviation being a one day overestimation with the MIROC model in the interval from 1980 to 2009. With the IPSL model, the bias is null in all periods, as with the MIROC model in the two future periods. The GFDL model, on the contrary, underestimates by less than half a day in the past and near future, while in the distant future this bias becomes an underestimation.

Dry annual maximum spells, as shown in Figure VIF, are a major concern: the three models greatly overestimate WRF results. In this case, the "best" model to use is the IPSL one, which underestimates this annual period by just over a month. This value is maintained in the three time periods. The GFDL model shows a similar behaviour to this one, but with a deviation of two months. The worst option for this metric is the use of the MIROC model, which from 1980 to 2009 presents a deviation of one and a half months, while in future periods this deviation soars to almost 3 months (80 days). This can be translated into a prediction of longer and drier summers.

Now, going back to ERA-5 results with the CNN model, the wet spells are overestimated by almost 3 days, whilst the dry ones are underestimated by almost a week. The difference with the values mentioned in the previous paragraph is clear: compared to the global models, the use of reanalysis data causes more difficulties in predicting wet spells, whereas it is, by far, more accurate in predicting dry ones.

Taking into account the values of the 5 metrics considered (the bias with respect to the 98th percentile was disregarded), the use of the IPSL model is the one that gives the best results (numerically speaking) in the three periods considered. However, due to the standard deviation of the predictions with this model, it is the least stable of the three with respect to the bias of the mean. However, it presents very good results if the RMSE of the wet days and the prediction of the maximum annual wet period are taken into account, and in spite of the not so favourable results, it is the best option for obtaining the maximum annual dry period. In the past and near future, the GFDL model is the worst overall performer, but in the far future, despite having very close values, this position is covered by the MIROC model.

As mentioned in the discussion of temperature results, global models data in the RCP 8.5 scenario is being used. Under these conditions, a decrease in precipitation is expected in the Canary Islands region, so it would be desirable that this trend would be mirrored in the predictions made by the statistical models developed in this work with the data from the global models. This is depicted in Figure VII.

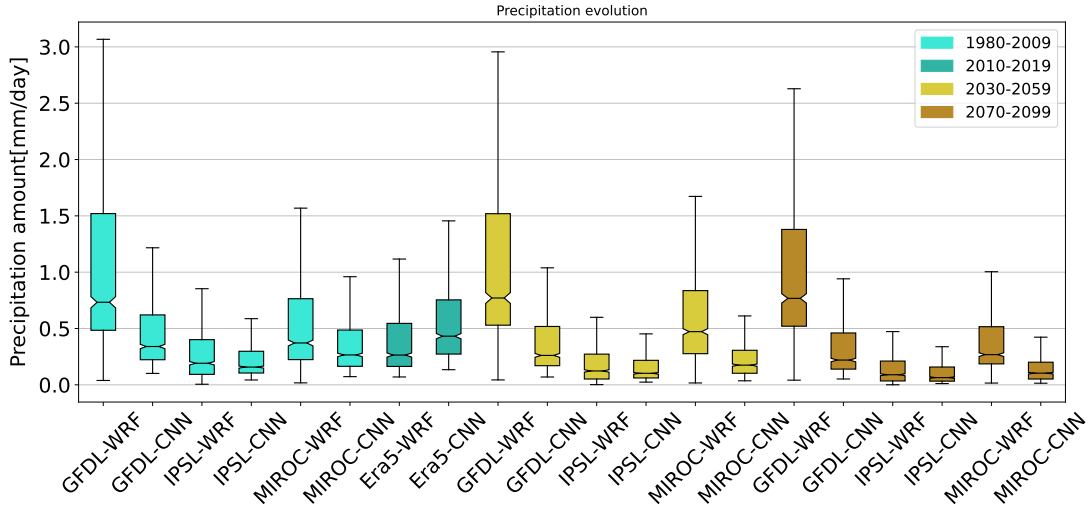
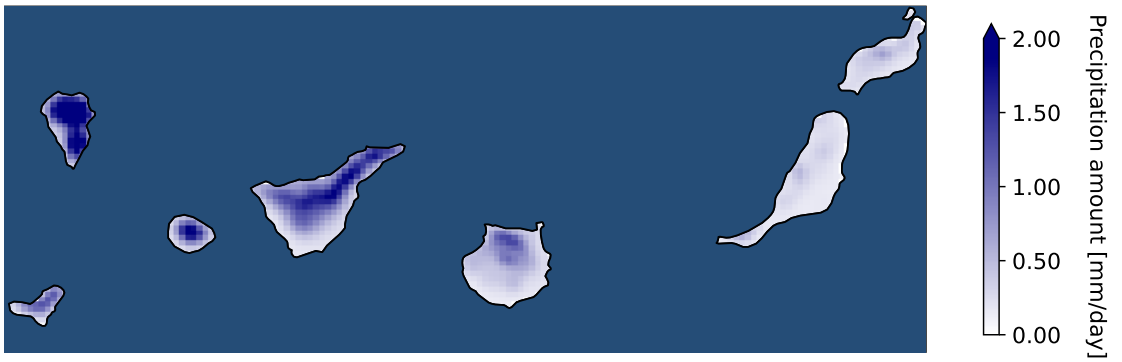


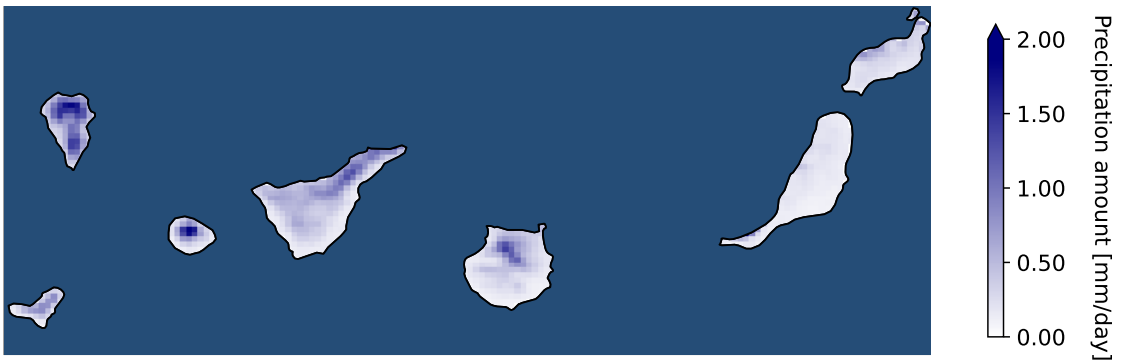
FIGURE VII: Predictions on precipitation amount for each model considered in this work. Results for CNN and WRF models are shown for four time periods: 1980-2009, 2010-2019, 2030-2059 and 2070-2099. The time period between 2010-2019 is only predicted with ERA-5 data, while the others are obtained from the GCMs models considered.

The first point to highlight is the range of values in which the plot moves: in the Canary Islands, currently (and in the near past) a mean precipitation of less than 1 mm/day is expected (De Luque Söllheim et al., 2023), thus the values obtained with the global models for the period from 1980 to 2009 are consistent with this known data. Due to these values, the downward trend in precipitation is not as clear as in the case of temperatures, but it is noticeable how each model predicts a lower value as time goes on. Under the RCP 8.5 scenario, a decrease in precipitation of about 20% is expected. GFDL results, as displayed in Figure VII, show a slight decrease over the years. The downward trend is better exposed by the MIROC model, but IPSL results are also consistent with the expected drop in precipitation.

The difference between the results of WRF and CNN is evident in Figure VII: those of the latter are always lower. The greatest difference is with the GFDL model, while the closest predictions are obtained using the IPSL model. Nonetheless, to a greater or lesser extent, all CNN results show a lower variability over the three periods considered. This is a problem in the predictions of extreme phenomena, and also leads to a general underestimation of precipitation.



(A) Predictions of WRF with MIROC data.

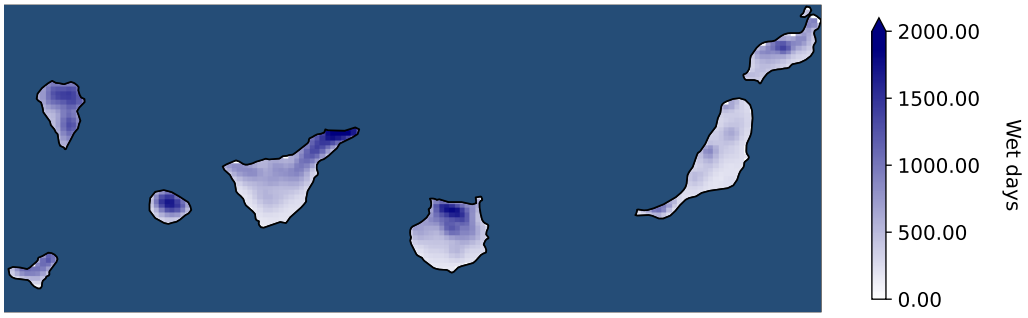


(B) Predictions of CNN with MIROC data.

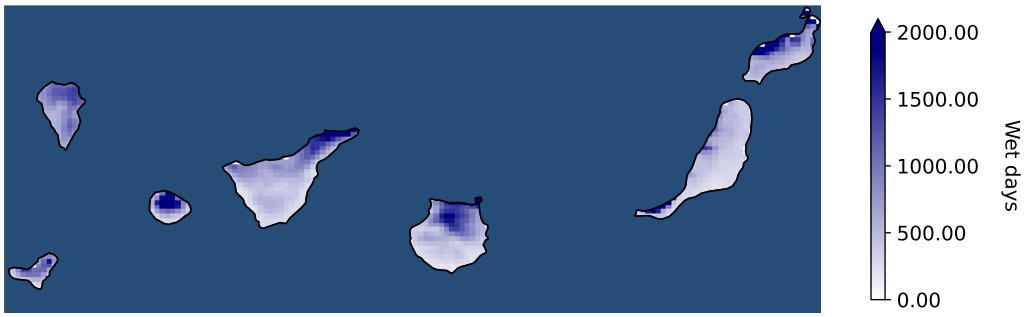
FIGURE VIII: Predictions on daily mean precipitation amount on the Canary Islands over 1980-2009 period. Comparison between WRF and CNN models.

Another noteworthy point is that the areas of the Canary Islands where it rains regularly and abundantly are limited: this includes *La Palma* (especially *La Caldera*), high elevation areas and the north of *Tenerife*, including the rural park of *Anaga* and the *Teide* National Park, and the area of the *Garajonay* National Park, in *La Gomera*. Areas such as the summits of *El Hierro* and *Gran Canaria* show less regularity and abundance, and the rest of the archipelago shows very scarce and infrequent precipitation. Due to this fact and the low overall precipitation amount, it is logical to think that the decrease in this magnitude will come hand in hand with a decrease in the abundance and/or frequency of precipitation in the aforementioned regions.

The Figure VIII displays the mean precipitation amount predictions in the Canary Islands for the period from 1980 to 2009. The problem commented above is evidenced here. The possible underlying reasons for the underestimation of precipitation by the CNN model are explored below, by comparing the intensity of precipitation on wet days and the number of wet days predicted by WRF and CNN. To this end, the results using the MIROC model will be used, as it is the closest to observations according to WRF predictions. However, these findings can be replicated for any model and period considered, as the underestimation of precipitation is a common problem for all of them.

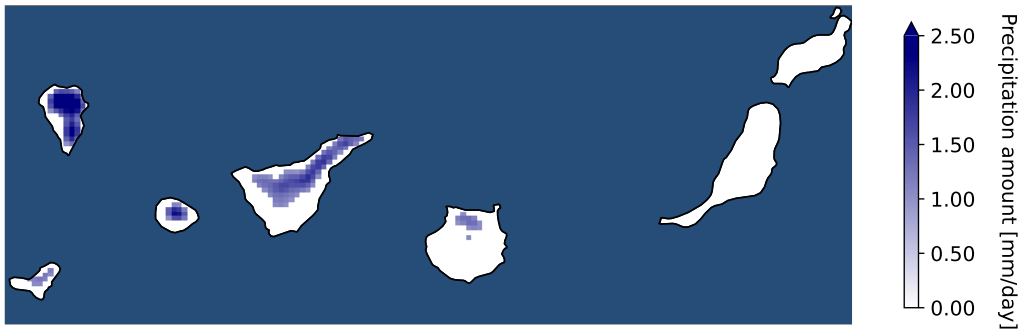


(A) Predictions of WRF with MIROC data.

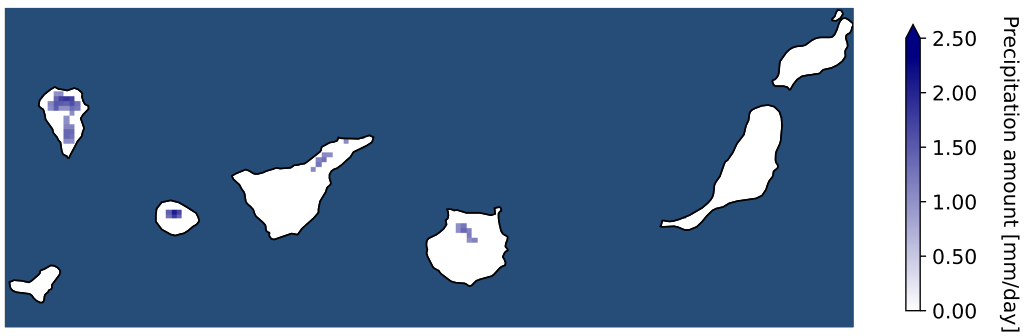


(B) Predictions of CNN with MIROC data.

FIGURE IX: Predictions on the amount of wet days on the Canary Islands over 1980-2009 period. Comparison between WRF and CNN models.



(A) Predictions of WRF with MIROC data.



(B) Predictions of CNN with MIROC data.

FIGURE X: Rain intensity predictions on the Canary Islands over 1980-2009 period (days above 1 mm/day). Comparison between WRF and CNN models.

When looking at the number of wet days (that is to say, days with a precipitation greater than 1 mm/day), it is striking to see in the Figure IX that the CNN (Figure IXB) predicts a higher number than WRF (Figure IXA) during the same time. The CNN shows a higher number of wet days with respect to WRF mainly in the north of the islands, with *La Gomera* and *Lanzarote* being particularly affected. These results suggest that the underestimation of precipitation is due to a much lower intensity of rain on wet days, as there is a difference of more than twice the amount of wet days.

In Figure X the intensity of precipitation on wet days is shown. Here, the shortcomings of the CNN with respect to WRF are evident. Furthermore, as seen due to the scarce variability of the CNN results in Figure VII, extreme precipitation events are also underestimated: this can be seen by comparing *La Palma* in both predictions (Figures XA and XB). Thus, according to the predictions of the CNN with MIROC in the period from 1980 to 2009, it rains less in the wettest areas of the Canary Islands: mainly *La Palma*, *La Gomera* and the north of *Tenerife*.

4 Conclusion

Resumen

Este último capítulo pone de manifiesto las conclusiones que se obtienen con base en los resultados previamente justificados.

The aim of this project was to develop two statistical downscaling CNN models for the Canary Islands, one for temperature and one for precipitation, capable of emulating the results of a dynamic downscaling, namely that of the GOTA WRF model. Let's review the suitability of these models in terms of the results obtained.

Regarding the results of mean temperature, the model developed performed well overall. Each dataset of each global model used had certain strengths and weaknesses: with the GFDL model, very low biases were obtained, as well as a reliable prediction of cold spells, while with the MIROC model, a good correlation between the variables was preserved, along with a low RMSE and dispersion, combined with a high accuracy in the duration of warm spells.

The precipitation model gives somewhat inadequate results, but similar to previous studies for the Canary Islands. The biases with respect to the mean are relatively large: as they are calculated as a percentage, this problem is caused by the driest areas, where a slight variation in the amount of precipitation results in a high percentage bias. This fact is in consonance with the high bias with respect to the annual maximum dry spells.

Overall, the problem with this model, derived from the analysis of the VALUE metrics and the predicted rainfall amount, is that the variability of precipitation is significantly reduced. It was concluded that this problem came from a poor prediction of intensity (lower than it should be) rather than an underestimation in the number of wet days.

The unsuitability of this CNN model structure for a statistical downscaling in the Canary Islands, training the network with ERA-5 reanalysis data, may suggest that some of the conditions described in section 1.1.1, on which the quality of the model is said to depend on, are not met. It is likely that the predictors used in this work are not sufficiently informative to describe the local variability of precipitations. In the wettest areas of the archipelago, this predictand shows an extremely high variability, both spatial and temporally. This hinders the prediction of precipitation both in dynamic and statistical models.

Another potential reason for this unsuitability could be that the meteorological processes considered require the underlying physical processes to be accurately predicted. This could explain why the results of the statistical model for precipitation is not as close to the dynamic model as the ones with the temperature model.

A potential improvement would be to train the models with past and present data. This could provide the models with more data variability and give them the possibility to better learn the patterns in the synoptic situation that lead to the generation of precipitation. In this framework, another cross-validation strategy, such as k-fold, could be implemented to better picture how the model will perform with new data.

A Cross-Entropy and Mean Square Error

A.1 Maximum Likelihood Estimation

The most common principle from which specific functions that are good estimators for different models can be derived is the maximum likelihood principle. Let be a set of n examples, $\mathbb{X} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}$. This set must be independent from the true data-generating distribution $p_{\text{data}}(\mathbf{x})$, which is unknown.

Now, let be $p_{\text{model}}(\mathbf{x}; \theta)$ a parametric family of probability distributions over the same space, indexed by θ. This family maps any configuration $\mathbf{x}$ to a real number, estimating the true probability $p_{\text{data}}(\mathbf{x})$.

Under these conditions, the maximum likelihood estimator for θ is defined as

$$\theta_{\text{ML}} = \arg\max_{\theta} p_{\text{model}}(\mathbb{X}; \theta) = \arg\max_{\theta} \prod_{i=1}^{n} p_{\text{model}}(\mathbf{x}^{(i)}; \theta) \tag{A.1}$$

Taking the logarithm of the likelihood does not change its argmax, but gives a more convenient optimization problem:

$$\theta_{\text{ML}} = \arg\max_{\theta} \sum_{i=1}^{n} \log p_{\text{model}}(\mathbf{x}^{(i)}; \theta) \tag{A.2}$$

Maximum likelihood estimation can be interpreted as minimizing the dissimilarity between the empirical distribution of the training set, $\hat{p}_{\text{data}}$, and the model distribution. The degree of dissimilarity between the two is measured by the Kullback-Leibler divergence, also called relative entropy. This divergence is a type of statistical distance that measures how one probability distribution diverges from a second one. The KL divergence is defined as

$$D_{KL}(\hat{p}_{\text{data}} || p_{\text{model}}) = E_{\mathbf{x} \sim \hat{p}_{\text{data}}}[\log \hat{p}_{\text{data}}(\mathbf{x}) - \log p_{\text{model}}(\mathbf{x})] \tag{A.3}$$

That way, when the two distributions are completely similar, the KL divergence will be zero. When training the neural network, the goal is to minimize this expression. As the logarithm of $\hat{p}_{\text{data}}$ cannot be changed, because it is only function of the data-generating process, the only thing that can be modified is the p_{model}. Therefore, the logarithm of the likelihood of the model is minimized. This means that when training the model, the only thing that is minimized is the cross-entropy between the two distributions. The cross-entropy is defined as

$$-E_{\mathbf{x} \sim \hat{p}_{\text{data}}}[\log \hat{p}_{\text{data}}(\mathbf{x})] \tag{A.4}$$

"Any loss function that can be written as the negative log-likelihood of the model is a cross-entropy between the empirical distribution defined by the training set and the probability distribution defined by the model", Goodfellow et al., 2016.

A.2 Mean Square Error as a Cross-Entropy

To show how the mean square error (MSE) relates to cross-entropy, linear regression from the maximum likelihood estimation point of view will be considered. Let be a model that produces a conditional distribution $p(y|\ \mathbf{x})$. Let be a learning algorithm with the goal of fitting the distribution $p(y|\mathbf{x})$ to all different values of y compatible with $\mathbf{x}$. Being this a linear regression problem, and because of the central limit theorem, it can be assumed that the distribution to be found is a Gaussian (its parameters: mean and variance). It is then defined

$$p(y|\mathbf{x}) = \mathcal{N}(y; \hat{y}(\mathbf{x}; \mathbf{w}), \sigma^2) \tag{A.5}$$

where the variance is fixed to σ^2 and the function $\hat{y}(\mathbf{x}; \mathbf{w})$ gives the predictions on the mean of the Gaussian by taking the input variable $\mathbf{x}$ and the weights $\mathbf{w}$ learned during the training. $\mathcal{N}$ is just showing that this is a Gaussian distribution. It can be demonstrated that the conditional log-likelihood is given by

$$\sum_{i=1}^{n} \log p(y^{(i)}|\mathbf{x}^{(i)}; \theta) = -n \log \sigma - \frac{n}{2} \log 2\pi - \sum_{i=1}^{n} \frac{||\hat{y}^{(i)} - y^{(i)}||^2}{2\sigma^2} \tag{A.6}$$

where $\hat{y}^{(i)}$ is the output of the linear regression on the i-th input $\mathbf{x}^{(i)}$ and n is the number of training samples. The MSE is defined as

$$\text{MSE}_{\text{train}} = \frac{1}{m} \sum_{i=1}^{m} ||\hat{y}^{(i)} - y^{(i)}||^2 \tag{A.7}$$

Comparing this with the log-likelihood, it can be seen that maximizing the latter with respect to $\mathbf{w}$ is equivalent to minimizing the MSE, both yield to the same estimate of the parameter $\mathbf{w}$. While the two criteria have different values, they have the same location of the optimum, so the use of MSE as a maximum likelihood estimation is justified.

B Data tables

This appendix contains some tables with the numerical results of the statistical downscaling models, shown on figures III and VI.

B.1 Temperature

	GFDL		IPSL		MIROC	
	Median	Mean ± std	Median	Mean ± std	Median	Mean ± std
Mean Bias	−0.9393	−0.7294 ± 1.0187	0.7387	0.6810 ± 0.8011	−0.1083	0.0082 ± 0.8521
P2 Bias	−0.6001	−0.7958 ± 1.1360	1.6698	1.4055 ± 1.2286	0.6262	0.3964 ± 1.1018
P98 Bias	−1.3622	−0.9827 ± 1.9888	−0.8765	−0.8192 ± 1.7700	−2.1761	−2.1575 ± 2.3219
R (Pearson)	0.6121	0.6053 ± 0.0555	0.4777	0.4786 ± 0.0841	0.8956	0.8797 ± 0.0535
std Ratio	0.9636	0.9979 ± 0.1802	0.8226	0.8414 ± 0.1473	0.8616	0.8721 ± 0.1723
RMSE	3.3805	3.5239 ± 0.7306	3.5241	3.5880 ± 0.8207	2.0778	2.1269 ± 0.3736
WAMS Bias	4.2000	4.2036 ± 1.2442	2.2000	2.1199 ± 1.0966	0.6000	0.5203 ± 1.2075
CAMS Bias	0.0000	0.0988 ± 1.1024	0.4000	0.4519 ± 0.7345	−0.4000	−0.4757 ± 0.7771

TABLE B.1: Temperature results from 1980 to 2009.

	GFDL		IPSL		MIROC	
	Median	Mean ± std	Median	Mean ± std	Median	Mean ± std
Mean Bias	−0.7941	−0.5684 ± 1.1199	1.2175	1.1806 ± 0.7915	−1.6549	−1.3997 ± 1.1815
P2 Bias	−0.4079	−0.5774 ± 1.0396	1.9438	1.7278 ± 1.1510	−1.3668	−1.4736 ± 1.0057
P98 Bias	−1.0248	−0.6102 ± 2.0495	−0.4906	−0.4609 ± 1.7969	−2.8238	−2.6266 ± 2.4110
R (Pearson)	0.4972	0.4913 ± 0.0630	0.3555	0.3548 ± 0.0822	0.8557	0.8481 ± 0.0399
std Ratio	0.9791	1.0054 ± 0.1617	0.7165	0.7528 ± 0.1470	0.9905	1.0020 ± 0.1785
RMSE	3.9147	4.0717 ± 0.8049	4.4672	4.5486 ± 0.8080	2.7996	2.8036 ± 0.4778
WAMS Bias	2.6000	2.5275 ± 1.1000	1.8000	1.9014 ± 1.2435	0.4000	0.3390 ± 0.9946
CAMS Bias	0.0000	−0.2614 ± 1.0490	0.6000	0.7388 ± 0.8298	0.2000	0.3381 ± 0.6898

TABLE B.2: Temperature results from 2030 to 2059.

	GFDL		IPSL		MIROC	
	Median	Mean ± std	Median	Mean ± std	Median	Mean ± std
Mean Bias	−0.7584	−0.5183 ± 1.1708	1.1135	1.0638 ± 0.7920	−2.0206	−1.7482 ± 1.2449
P2 Bias	−0.2460	−0.3914 ± 1.0083	1.8156	1.6048 ± 1.1142	−1.6046	1.6086 ± 1.0914
P98 Bias	−1.6199	−1.3116 ± 2.0512	−0.1458	−0.1361 ± 1.5341	−2.8974	−2.9301 ± 2.2210
R (Pearson)	0.4135	0.4078 ± 0.0596	0.2900	0.2904 ± 0.0819	0.9111	0.8963 ± 0.0494
std Ratio	0.9639	0.9757 ± 0.1381	0.7591	0.7858 ± 0.1297	0.9849	0.9760 ± 0.1441
RMSE	4.4705	4.6219 ± 0.9138	4.5873	4.7069 ± 0.8868	2.7612	2.8003 ± 0.6538
WAMS Bias	1.0000	0.9333 ± 1.2471	1.6000	1.5016 ± 1.2271	0.8000	0.9122 ± 1.0406
CAMS Bias	−0.2000	−0.2276 ± 1.3091	0.8000	0.8251 ± 0.7988	−0.4000	−0.4693 ± 0.8828

TABLE B.3: Temperature results from 2070 to 2099.

B.2 ERA-5

	Temperature	
	Median	Mean ± std
Mean Bias	0.8217	0.8225 ± 0.0765
P2 Bias	1.1757	1.1934 ± 0.2866
P98 Bias	-0.2968	-0.3281 ± 0.5781
R (Pearson)	0.9048	0.8963 ± 0.0544
std Ratio	0.9199	0.9137 ± 0.0458
RMSE	1.6333	1.6538 ± 0.1505
WAMS Bias	1.4000	1.5401 ± 1.2392
CAMS Bias	1.4000	1.3866 ± 1.8709

TABLE B.4: Temperature results from 2010 to 2019 (ERA-5 data).

	Precipitation	
	Median	Mean ± std
Mean Bias	49.9561	54.6117 ± 36.2191
P98 Bias	81.3842	-
R (Spearman)	0.4664	0.4677 ± 0.0982
RMSE (Wet Days)	2.2590	2.7074 ± 1.4598
Wet AMS Bias	2.6	3.5615 ± 4.4052
Dry AMS Bias	-6.8	-10.6999 ± 21.6964

TABLE B.5: Precipitation results from 2010 to 2019 (ERA-5 data).

B.3 Precipitation

	GFDL		IPSL		MIROC	
	Median	Mean ± std	Median	Mean ± std	Median	Mean ± std
Mean Bias	-57.6576	-50.1047 ± 36.64	-10.5072	17.5285 ± 186.6801	-32.4197	-18.3049 ± 74.7331
P98 Bias	-44.8874	-	-	-	-	-
R (Spearman)	0.0044	0.0055 ± 0.01390	0.0111	0.01490 ± 0.02640	0.3564	0.3445 ± 0.07220
RMSE (wet days)	5.2756	6.9683 ± 4.2819	1.7815	2.4974 ± 1.7275	3.0861	4.1364 ± 2.6067
Wet AMS Bias	0.4000	1.0956 ± 3.2181	0.0000	1.0990 ± 5.7995	1.0000	1.9836 ± 4.4420
Dry AMS Bias	66.2000	60.9284 ± 38.3877	35.2000	41.4699 ± 53.2004	44.4000	36.5749 ± 42.6571

TABLE B.6: Precipitation results from 1980 to 2009.

	GFDL		IPSL		MIROC	
	Median	Mean ± std	Median	Mean ± std	Median	Mean ± std
Mean Bias	-68.3302	-61.0311 ± 30.4247	-5.8424	41.5628 ± 370.6872	-64.54	-51.0909 ± 70.5689
P98 Bias	-	-	-	-	-	-
R (Spearman)	0.01350	0.01560 ± 0.01820	0.01340	0.01720 ± 0.02450	0.2478	0.2495 ± 0.06630
RMSE (wet days)	10.0435	10.9919 ± 4.4610	1.4840	1.9874 ± 1.4319	3.0772	3.7460 ± 2.0674
Wet AMS Bias	0.2000	1.0130 ± 3.9940	0.0000	1.8402 ± 8.8521	0.0000	1.3432 ± 6.2422
Dry AMS Bias	69.4000	60.2240 ± 41.7419	34.0000	39.7841 ± 54.9484	80.6000	68.7603 ± 51.9128

TABLE B.7: Precipitation results from 2030 to 2059.

	GFDL		IPSL		MIROC	
	Median	Mean ± std	Median	Mean ± std	Median	Mean ± std
Mean Bias	-72.6201	-64.4522 ± 33.8718	-19.4288	110.2007 ± 1571.1628	-69.0458	-45.8048 ± 93.7937
P98 Bias	-	-	-	-	-	-
R (Spearman)	0.0123	0.0116 ± 0.0132	0.0019	0.0056 ± 0.0196	0.1931	0.2007 ± 0.0612
RMSE (wet days)	10.2717	11.1503 ± 4.2548	1.0937	1.4941 ± 1.1650	4.3645	5.1262 ± 2.5213
Wet AMS Bias	-0.2000	1.1228 ± 5.2282	0.0000	2.7116 ± 12.2501	0.0000	1.6149 ± 7.4986
Dry AMS Bias	70.4000	59.5994 ± 43.8780	33.0000	42.7534 ± 59.6124	81.8000	69.4948 ± 62.0040

TABLE B.8: Precipitation results from 2070-2099.

References

Ashraf, Rehan, Muhammad Asif Habib, Muhammad Akram, Muhammad Ahsan Latif, Muhammad Sheraz Arshad Malik, Muhammad Awais, Saadat Hanif Dar, Toqeer Mahmood, Muhammad Yasir, and Zahoor Abbas (2020). "Deep Convolution Neural Network for Big Data Medical Image Classification". In: *IEEE Access* 8, pp. 105659–105670. ISSN: 21693536. DOI: 10.1109/ACCESS.2020.2998808.

Baño Medina, Jorge (2021). "Deep Convolutional neural networks for statistical downscaling of climate change projections". In: *PhD Thesis. Universidad de Cantabria*.

Baño Medina, Jorge, Rodrigo Manzanas, Ezequiel Cimadevilla, Jesus Fernandez, Jose Gonzalez-Abad, Antonio S. Cofinõ, and Jose Manuel Gutierrez (Sept. 2022). "Downscaling multi-model climate projection ensembles with deep learning (DeepESD): Contribution to CORDEX EUR-44". In: *Geoscientific Model Development* 15 (17), pp. 6747–6758. ISSN: 19919603. DOI: 10.5194/gmd-15-6747-2022.

Baño Medina, Jorge, Rodrigo Manzanas, and José Manuel Gutiérrez (Apr. 2020). "Configuration and intercomparison of deep learning neural models for statistical downscaling". In: *Geoscientific Model Development* 13 (4), pp. 2109–2124. ISSN: 19919603. DOI: 10.5194/gmd-13-2109-2020.

Berrar, Daniel (Jan. 2018). "Cross-validation". In: vol. 1-3. Elsevier, pp. 542–545. ISBN: 9780128114322. DOI: 10.1016/B978-0-12-809633-8.20349-X.

Bony, Robert Colman, Vladimir M. Kattsov, Richard P. Allan, Christopher S. Bretherton, Jean-Louis Dufresne, Alex Hall, Stephane Hallegatte, Marika M. Holland, William Ingram, David A. Randall, Brian J. Soden, George Tselioudis, and Mark J. Webb (2006). "How Well Do We Understand and Evaluate Climate Change Feedback Processes?" In: *Journal of Climate* 19.15, pp. 3445 –3482. DOI: 10.1175/JCLI3819.1. URL: https://journals.ametsoc.org/view/journals/clim/19/15/jcli3819.1.xml.

Cannon, Alex J. (2008). "Probabilistic multisite precipitation downscaling by an expanded Bernoulli-gamma density network". In: *Journal of Hydrometeorology* 9 (6). ISSN: 1525755X. DOI: 10.1175/2008JHM960.1.

De Luque Söllheim, Angel Luis, Fabian García Hernández, and Pablo Lucas Mayer Suarez (2023). "The Digital Climate Atlas of the Canary Islands: a Tool to Improve Knowledge of Climate and Climate Change in the Canary Islands". In: *Available at SSRN 4470014*. DOI: https://dx.doi.org/10.2139/ssrn.4470014.

Ekström, Marie, Michael R Grose, and Penny H Whetton (2015). "An appraisal of downscaling methods used in climate change research". In: *WIREs Climate Change* 6.3, pp. 301–319. DOI: https://doi.org/10.1002/wcc.339.

Expósito, Francisco J, Albano González, Juan C Pérez, Juan P Díaz, and David Taima (2015). "High-resolution future projections of temperature and precipitation in the Canary Islands". In: *Journal of Climate* 28.19, pp. 7846–7856. DOI: https://doi.org/10.1175/JCLI-D-15-0030.1.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. http://www.deeplearningbook.org. MIT Press.

Guinea, Carlos Correa, Alfonso Hernanz Lázaro, and Esteban Rodríguez Guisado (2023). *Evaluación de métodos de regionalización estadística para la generación de proyecciones climáticas en el marco del PNACC-2 2021-2030*. Agencia Estatal de Meteorología. DOI: 10.31978/666-23-009-0.

Hersbach, Hans et al. (2020). "The ERA5 global reanalysis". In: *Quarterly Journal of the Royal Meteorological Society* 146.730, pp. 1999–2049. DOI: https://doi.org/10.1002/qj.3803.

Li, Kexuan, Fangfang Wang, Lingli Yang, and Ruiqi Liu (June 2023). "Deep feature screening: Feature selection for ultra high-dimensional data via deep neural networks". In: *Neurocomputing* 538. ISSN: 18728286. DOI: 10.1016/j.neucom.2023.03.047.

Maraun, Douglas and Martin Widmann (Jan. 2018). *Statistical Downscaling and Bias Correction for Climate Research*. Cambridge University Press, pp. 133–224. ISBN: 9781107066052. DOI: 10.1017/9781107588783.

Maraun, Douglas, Martin Widmann, José M. Gutiérrez, Sven Kotlarski, Richard E. Chandler, Elke Hertig, Joanna Wibig, Radan Huth, and Renate A.I. Wilcke (2015). "VALUE: A framework to validate downscaling approaches for climate change studies". In: *Earth's Future* 3.1, pp. 1–14. DOI: https://doi.org/10.1002/2014EF000259.

Medar, Ramesh, Vijay S. Rajpurohit, and B. Rashmi (2017). "Impact of Training and Testing Data Splits on Accuracy of Time Series Forecasting in Machine Learning". In: pp. 1–6. DOI: 10.1109/ICCUBEA.2017.8463779.

Pérez, JC, JP Díaz, A González, J Expósito, F Rivera-López, and D Taima (2014). "Evaluation of WRF parameterizations for dynamical downscaling in the Canary Islands". In: *Journal of climate* 27.14, pp. 5611–5631. DOI: https://doi.org/10.1175/JCLI-D-13-00458.1.

Perez, Juan C, Francisco J Exposito, Albano Gonzalez, and Juan P Diaz (2022). "Climate projections at a convection-permitting scale of extreme temperature indices for an archipelago with a complex microclimate structure". In: *Weather and Climate Extremes* 36, p. 100459. DOI: https://doi.org/10.1016/j.wace.2022.100459.

Reitermanová, Z. (2010). *Data Splitting*. ISBN: 9788073781392.

Riahi, Keywan, Shilpa Rao, Volker Krey, Cheolhung Cho, Vadim Chirkov, Guenther Fischer, Georg Kindermann, Nebojsa Nakicenovic, and Peter Rafaj (Nov. 2011). "RCP 8.5-A scenario of comparatively high greenhouse gas emissions". In: *Climatic Change* 109 (1), pp. 33–57. ISSN: 01650009. DOI: 10.1007/s10584-011-0149-y.

Skamarock, C., Bogumiła Klemp, Jimy Dudhia, O. Gill, Zhiquan Liu, Judith Berner, Wei Wang, G. Powers, Greg Duda, Dale Melvyn Barker, and Xiangyu Huang (2019). "A Description of the Advanced Research WRF Model Version 4". In: URL: https://api.semanticscholar.org/CorpusID:196211930.

Taylor, Karl E, Ronald J Stouffer, and Gerald A Meehl (2012). "An overview of CMIP5 and the experiment design". In: *B. Am. Meteorol. Soc.* 93.4, pp. 485–498.

Wilby, Robert, Christian Dawson, Conor Murphy, P. O'Connor, and E. Hawkins (Jan. 2014). "The Statistical Downscaling Model - Decision Centric (SDSM-DC): conceptual basis and applications". In.

Williams, Peter (Jan. 1997). "Modelling Seasonality and Trends in Daily Rainfall Data." In.