Anonymous Referee #1

In this manuscript, deep learning methodology based on CNN architectures is evaluated for downscaling CMIP6 simulations and projection with standard resolutions to a higher resolution, i.e. 1/10th deg. The analysis focuses on the Iberian Peninsula, and the ability of four CNN architectures to reproduce the surface T, Tmin, Tmax and Pr climates is evaluated. The DL algorithm is trained using ERA5 and compared to the high-resolution regular gridded dataset over the historical period, and then used to downscale future projections (relative to historical climate) in agreement with four future scenarios and multi-models.

Overall, the manuscript is well organized in format and the writing is clear. Although the physical process is rarely touched (due to the limitation of DL), the downscaled results based on DL are sound. I have some comments regarding clarifications in the main text.

We would like to thank the reviewer for their availability to read and comment our manuscript. We are grateful for your positive and constructive comments and suggestions. Our response to each comment is given below. We sincerely think that your revision allowed an overall improvement of the manuscript.

Line 23: "Notably, …. climate scenario". Please specify the region (Iberia) exhibiting the temperature increase.

We have now specified the region in the sentence: "*Notably, a clear warming trend is observed in Iberia*, … climate scenario*"*.

Line 129: I am not familiar with any paper using "democratic" to describe a simple average. It would be better to just use "simple average", which is straightforward and easy to follow. I noticed that "democratic" has been mentioned several times in this manuscript. Please also revise those instances accordingly.

We have implemented the reviewer's recommendation by incorporating the suggested changes throughout the entire text, using the terminology "simple average".

Line 145 (Fig. 1) Can you mention why this domain of predictors is chosen (dashed line)? Have you tested the sensitivity of the downscaled fields when the domain of predicators is changed?

The chosen domain for the predictors is larger than the output domain to ensure that the necessary information about large-scale phenomena given by the predictors is provided to perform a solid downscaling. We made an initial test with a smaller domain, Iberia-only, similar to the output domain. The downscaling results were overall slightly inferior in quality when using a smaller domain. To maintain computational efficiency, we did not try downscaling with larger domains.

We added the following sentence in line 148: *“The predictors domain (Fig. 1, red dashed line) is a larger region than the IP domain to ensure that large-scale phenomena are included in the information provided by the predictors to train the DL models.”*

Lines 170-173: These sentences would be better placed in the introduction.

The sentences were moved to the introduction (lines 113-116). We also added an introductory sentence about Iberia01 prior to those sentences (lines 111-113): “*The Iberia01 regular gridded product (hereafter Iberia01) is the highest resolution observational daily dataset including mean, maximum and minimum temperatures and precipitation, covering the full domain of continental Iberia (Herrera et al., 2019).”*

Table 1: It would be better to provide the full name of CMIP6 ESGCM. For example, CM6A-LR -> IPSL-CM6A-LR.

Full names provided.

Lines 220-222: I am not able to follow this sentence. Could you rephrase it?

We rephrased the sentence as follows: *“For precipitation, however, the DL models feature a multi-output structure (see Fig. 3 of Baño-Medina et al., 2020).”*

Lines 222-223: How are these terms (alphas, beta, and distribution) used in this study? I couldn’t find more discussions about these terms from CMIP6 runs.

These terms are the output of the DL models used to downscale precipitation. The precipitation value is then obtained by multiplying the alpha and beta terms. The results shown are discussed based on the resulting precipitation value. We added a new sentence in line *231: “The precipitation value is obtained by multiplying the alpha and beta parameters.”*

Line 238: How were the missing data filled? Were they filled with zeros?

The process of filling missing data in a grid point involved taking the average of the value from neighbouring grid points at the same time step. In cases where the surrounding grid points also had missing data, we resorted to using a domain-wide average for the replacement.

We added the following sentence to the manuscript in line 243: *“Gridpoints with missing data in the CMIP6 ESGCMs were filled with an average of the surrounding gridpoints. If the surrounding gridpoints had missing data as well, a domain average was applied. Afterwards, the dataset was standardized (with the same parameters used for ERA5).”*

Line 276: How about changing “members” to “models”, as the members are also used to describe 4 DL architectures (Line 384)? Similarly, in Line 48, 7 members -> 7 models.

Where appropriate, we replaced “members” with “models”.

Line 283: Why was the base period chosen as 1981-2010 and not 1979-2014? The projected temperature increase depends on the reference period, which should be mentioned clearly in the corresponding section.

1981-2010 was chosen because it represents a 30-year climatology that is commonly used as reference for the historical period in climate studies, and also as a climatological normal period for several institutions, such as the Portuguese and Spanish meteorological institutes. Also, this way, the base period and the future periods (except 2015-2040) consider the same 30-year length.

We added the following sentence in line 299: *"It should be noted that the projected temperature increase depends on the chosen historical period."*

Fig. 2 and subsequent figures: I am not sure how the error bar of each boxplot is calculated. Is it related to the uncertainty of parameters in the DL model?

In Figures 2-5, the bar in each boxplot represents the value of all gridpoints of the output of each CNN method forced with ERA5. In Figures 6-9, the bar in each boxplot represents all gridpoints of the output of all CNN methods pooled together forced with each GCM.

For clarification, we added a description of the boxplot in each figure legend.

The error bar in Fig. 10 and the following figures seem to have different meanings compared with the previous figures, I guess. Are these related to the spread from 4 CNN methods and 7 models?

The bar in each boxplot represents the value of all gridpoints of the output of all CNN methods forced by all GCMs, all pooled together.

As in the previous comment, we also added a description of the boxplot in each figure legend for clarification purposes.