We thank the reviewer for his/her fruitful comments and sincerely acknowledge his/her time for reviewing the manuscript.

- The sentence “To our knowledge, this is the first time that CNNs have been used to produce multi-model ensembles” is not that accurate since there are previous studies that employed CNN to downscale the model ensemble (e.g., Babaousmail et al. (2021)).

  We thank the reviewer for this interesting reference. We have noticed that in [1], the Deep Learning topology is used as a Model Output Statistics (MOS) technique, since it uses the precipitation —which is also the target variable— of the Generalized Circulation Model (GCM) as input data. This differs with our approach which falls in the “perfect-prognosis” family within statistical downscaling, using reanalysis data as input to calibrate the model. For this reason we have rephrased the sentence indicated by the reviewer to the following one: “To our knowledge, this is the first time that CNNs have been used to produce downscaled multi-model ensembles based on the perfect-prognosis approach”.

  .

  [1] Babaousmail, Hassen, et al. "Novel statistical downscaling emulator for precipitation projections using deep Convolutional Autoencoder over Northern Africa." *Journal of Atmospheric and Solar-Terrestrial Physics* 218 (2021): 105614.

- Introduction (third paragraph): “These methods are not computationally demanding…”. This sentence needs a citation.

  As compared with the Regional Climate Models (RCMs), which require computational timings from month to years in their simulations, statistical models including convolutional neural networks can be calibrated from hours to days. We have included the following citations:
  [1] Le Roux, Renan, et al. "Comparison of statistical and dynamical downscaling results from the WRF model." Environmental modelling & software 100 (2018): 67-73.
  [2] Baño-Medina, Jorge, Rodrigo Manzanas, and José Manuel Gutiérrez. "Configuration and intercomparison of deep learning neural models for statistical downscaling." Geoscientific Model Development 13.4 (2020): 2109-2124.

- The author should justify why he selected the RCP8.5 scenario out of the other scenarios?
  We have included the following sentence in the manuscript “We follow previous work in the field [1,2] and select the RCP8.5 scenario which shows the strongest climate change signal (especially for temperature) and, therefore, permits to optimally explore the extrapolation capability of the CNNs.”

[1] *ME, Olmo, Rocio Balmaceda-Huarte, and Maria Laura Bettolli. "Multi-model ensemble of statistically downscaled GCMs over southeastern South America: historical evaluation and future projections of daily precipitation with focus on extremes." Climate Dynamics (2022): 1-18.*
[2] *Baño-Medina, Jorge, Rodrigo Manzanas, and José Manuel Gutiérrez. "On the suitability of deep convolutional neural networks for continental-wide downscaling of climate change projections." Climate Dynamics 57.11 (2021): 2941-2951.*

- Also, was there any method employed for the selection of the 8 GCMs?
  The main reason for selecting this ensemble is because it has been already used in other studies so it allows comparison. Moreover, predictors are publicly available (allowing reproducibility, as illustrated in the companion notebook) and have been assessed in a previous study [1]..[1] *Brands, Swen, et al. "How well do CMIP5 Earth System Models simulate present climate conditions in Europe and Africa?." Climate dynamics 41.3 (2013): 803-817.*

- The author didn't justify why E-OBS v20 was selected as an observation in this study.
  We have added the following paragraph to the manuscript:
  "E-OBS is a high-resolution observational dataset generated through an interpolation procedure of the European Climate Assessment & Dataset (ECA&D, [1]) station network. Whilst national and sub-national datasets exist, E-OBS accurately represents the regional climate over the entire European continent [2] and it is commonly employed in statistical downscaling experiments at a continental level [3,4,5,6]. We chose version 20 (v20, release date October 2019) since it was the latest one at the beginning of this study."

  [1] Klok, E. J., and A. M. G. Klein Tank. "Updated and extended European dataset of daily climate observations." *International Journal of Climatology: A Journal of the Royal Meteorological Society* 29.8 (2009): 1182-1191.
  [2] *Bandhauer, Moritz, et al. "Evaluation of daily precipitation analyses in E-OBS (v19. 0e) and ERA5 by comparison to regional high-resolution datasets in European regions." International Journal of Climatology 42.2 (2022): 727-747.*
  [3] *Maraun, Douglas, et al. "VALUE: A framework to validate downscaling approaches for climate change studies." Earth's Future 3.1 (2015): 1-14.*
  [4] *Vrac, Mathieu, and Pradeebane Vaittinada Ayar. "Influence of bias correcting predictors on statistical downscaling models." Journal of Applied Meteorology and Climatology 56.1 (2017): 5-26.*
  [5] *Baño-Medina, Jorge, Rodrigo Manzanas, and José Manuel Gutiérrez. "Configuration and intercomparison of deep learning neural models for statistical downscaling." Geoscientific Model Development 13.4 (2020): 2109-2124.*
  [6] *Baño-Medina, Jorge, Rodrigo Manzanas, and José Manuel Gutiérrez. "On the suitability of deep convolutional neural networks for continental-wide downscaling of climate change projections." Climate Dynamics 57.11 (2021): 2941-2951.*

- Since the author is comparing the ensemble resulting from 8 GCMs with the RCM ensemble projection, shouldn't be the number of GCMs equal to the number of RCMs?
  As the reviewer mentions, ideally we would have the same number of members in both GCM and RCM ensembles. Under this assumption there would only be 1 RCM

per GCM resulting in a total of 8 members for both ensembles. However, we wanted to avoid possible artifacts in the results due to a lack of variability in the RCM selection. For this reason we occasionally utilize 2 RCMs per GCM, representing a compromise between having a similar —but not equal— number of members for each ensemble and partially including this source of uncertainty in the results.

- Usually, when we train a neural net model, a validation phase is required after the training and it should be selected from the historical 25 years period, in this paper the author didn't mention it.
  During training, we use a validation set (10% of the data) to cross-validate the results performing early-stopping [1] (the training stops when the validation test start increasing). We have included the following phrase in the annex of the manuscript to mention this aspect: "During calibration, we use a validation set (10% of the data randomly selected) to perform early-stopping [1], and finish the training whenever the validation loss stops decreasing after 30 epochs" . After training, the validation in ''perfect" conditions was already carried out in [1] and it is appropriately cited in lines 40 and 74-75.

  [1] *Baño-Medina, Jorge, Rodrigo Manzanas, and José Manuel Gutiérrez. "Configuration and intercomparison of deep learning neural models for statistical downscaling." Geoscientific Model Development 13.4 (2020): 2109-2124.*

- Concerning the CNN algorithm, we noticed that the CNN used to downscale precipitation has one more layer than the one for temperature (one output layer). Can the author explain the reason?
  The configuration of the DeepESD method was undertaken in [1] and here we used the optimum configurations found in that study. To downscale both temperature and precipitation fields, the CNNs deployed contain the same number of hidden layers (3) but different number of, n, output neurons (2*n for temperature and 3*n for precipitation; please note that output layers are arranged in parallel and not sequentially), representing each of the statistical parameters of the parametrics distributions estimated per gridbox. For temperature, we learn Gaussian daily conditional distributions parameterized by 2 parameters (mean and standard deviation) per predictand site, while for precipitation there are 3 parameters (probability of rain, shape and scale factor) corresponding to the Bernoulli-Gamma distribution. This results in 2 and 3 output layers in the convolutional topology, respectively. More details can be found in the reference describing the convolutional network here employed [1], already cited in the manuscript.

  [1] *Baño-Medina, Jorge, Rodrigo Manzanas, and José Manuel Gutiérrez. "Configuration and intercomparison of deep learning neural models for statistical downscaling." Geoscientific Model Development 13.4 (2020): 2109-2124.*