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

- As RCMs require massive computational resources, the CNN-based model may be plausible to be used as an alternative for downscaling GCM outputs. So, please add some discussions about the computational efficiency of the CNN-based model.
  We have included the following sentence: “As compared to DD, PP lacks explicit physics in the model formulation, but overcomes systematic biases present in RCM products, since the model is trained using observations. Regarding computational requirements PP has smaller requirements avoiding the need for large computational infrastructures [1,2]. These aspects make PP models attractive to be extensively used to downscale global multi-model ensembles providing continental-wide regional projection fields, e.g. over the CORDEX domains, a key task which is mostly undertaken by means of DD nowadays”

  [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.

  We have also added the computational resources and timings required for training and prediction of our particular use-case in the manuscript: “We lean on a 2x Intel(R) Xeon(R) CPU E5-2670 0 @ 2.60GHz (16 cores, 32 threads) with 60 GB of memory RAM. The computational time employed to calibrate the model and generate the projections for e.g., one GCM is approximately 5-6 hours, being considerably lower than the one needed to run a RCM (weeks to months)”
  Moreover, we cite [2] which compares the computational efficiency of the same CNN topology here deployed against other statistical models.

  [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.

- Line 12: “… but a similar uncertainty for temperature”. More precisely, it is a larger uncertainty for temperature, as shown in Figure 3.
  Yes, thank you for noticing it. We have rephrased it to “... but a larger uncertainty for temperature”.

- Line 71−73: The CNN-based model provides smaller uncertainties for precipitation. Does the harmonization process contribute to this respect? Please add discussions about the role of the harmonization process.
  In the historical period the predictor fields are all harmonized (bias adjusting the mean and variance of the seasonal cycle) with ERA-Interim —regardless of the GCM considered,— thus leading to a small ensemble spread. Nevertheless, in the future the ensemble spread of DeepESD is the result of 1) the divergence in the

trends/evolution of the predictor fields, and/or to 2) the extrapolation ability of the CNN. Nevertheless, we agree with the reviewer that the harmonization may alter the trends in the projected signals [1] and may have an impact on the inter-model uncertainty. We are currently exploring the influence of different bias adjustment techniques over the predictor space in the downscaled signal, and therefore a more extensive analysis on this task will be published as future work. In the meantime we have added these two sentences in the manuscript:
(Data and Methods) "Since advanced bias adjustment techniques may alter the raw climate change signal [1], we only adjust the mean and variance at a monthly scale to keep this as simple as possible."
(conclusions) "Also, the sensitivity of the projected signals to the bias adjustment of the predictor fields is to be explored in future work."

[1] *Casanueva, Ana, et al. "Testing bias adjustment methods for regional climate change applications under observational uncertainty and resolution mismatch." Atmospheric Science Letters 21.7 (2020): e978.*

- Figure 1. Please add units (°C) to the color bar of the temperature bias.
  Done. Thank you for noticing it.

- Line 102. The CNN-based model is trained with the observations. So it can be expected that DeepESD exhibits a largely unbiased spatial pattern. The CNN-based model is also used to downscale the GCM outputs for the historical period. How about the downscaled historical GCM simulations by the CNN-based model? Please add discussions about the bias of the CNN-based model for the historical period.
  Indeed, the bias pattern displayed in Figure 1, row 2, columns 3 and 6, is the climatological difference of the downscaled historical mean of precipitation and temperature obtained with DeepESD, and the same field for the raw GCM. So this figure is already the one demanded by the reviewer and is discussed in the manuscript. We have clarified this aspect by adding the following sentence: " The bottom row displays the corresponding biases of the raw (GCM) and downscaled historical simulations (RCM and DeepESD) with respect to E-OBS v20."

  Moreover, as shown in [1], the bias of the downscaled mean of temperature and precipitation for the same reference period using ERA-Interim predictors as input data, shows also a mostly unbiased pattern.

  [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.*

- L106. Figure 2 shows that DeepESD shows good performance in reproducing the variability and extremes. I think this capability is important. Please add more discussions about Figure 2.
  We have added the following paragraph in the manuscript:
  "Besides these results for the mean, Figure 2 compares the entire precipitation (>= 1 mm/day) and temperature distributions for the GCM (red), RCM (blue) and DeepESD

(green) ensembles over the historical period 1979-2005, plus E-OBS (black), for the Alps, Iberian Peninsula and Eastern Europe as illustrative for different European climates. The solid line represents the ensemble mean and the shadow encompasses two standard deviations. The dashed line indicates the distributional mean of each PDF. For precipitation, we observe a good fit between E-OBS and DeepESD. This is a direct consequence of sampling from the inferred conditional Gamma distributions. RCMs and, particularly, GCMs overestimate low rainfall events and underestimate the high rainfall ones. Although the differences among the three ensembles for temperature are not as notable as for precipitation, DeepESD tends to follow more accurately the E-OBS curve than the GCM and the RCM."

[1] *Carreau, Julie, and Mathieu Vrac. "Stochastic downscaling of precipitation with neural network conditional mixture models." Water Resources Research 47.10 (2011).*

- L114−116: I do not quite understand this sentence. Could you please explain it?
  We agree with the reviewer that the sentence might be confusing. We have rephrased it to: "However, slight regional differences exist among the three ensembles, especially when compared with GCMs, with DeepESD presenting weaker signals of change over the Iberian Peninsula and more intense signals over some parts of Northern and Eastern Europe."