MAJOR COMMENTS:

1. It would be good to say something more specifically about European applications and the related data needs. Climate change studies are mentioned in the text. But nothing is said about the types of applications/users in the Introduction –and this influences the types of information required – e.g., whether spatial consistency is important, the types of extremes that are relevant.

Sectoral studies, such as hydrology, agriculture or energy applications, are in need of high-resolution climate/meteorological information at different time scales. For example at short scales (i.e, hourly and daily), accurate forecasts of wind fields are crucial to predict the capacity of renewable sources to meet the demands of the energy market. At longer scales, the impact and adaptation communities derive indices from the downscaled climate projections to evaluate the influence of climate change in different environments (e.g., health, agriculture). In addition, certain applications are sensitive to the spatial consistency of the downscaled information (e.g., to evaluate the impacts on water resources over a certain area) or to their suitability to accurately reproduce extremes (e.g., droughts and floods can cause devastating damages in agriculture).

We have included the above paragraph in the new version of the manuscript.

2. Line 33-39: This study focuses on the deep learning techniques in the context of perfect prognosis SD. However, it's not clear what the difference between classical SD methods and machine learning techniques is. This needs to be mentioned in the introduction.

By classical statistical downscaling techniques we refer to traditional and well established approaches adopted by the climate community, including generalized linear models, analogs and model output statistics, but also bias correction. In the machine learning paradigm there are more sophisticated approaches such as random forests, neural networks and

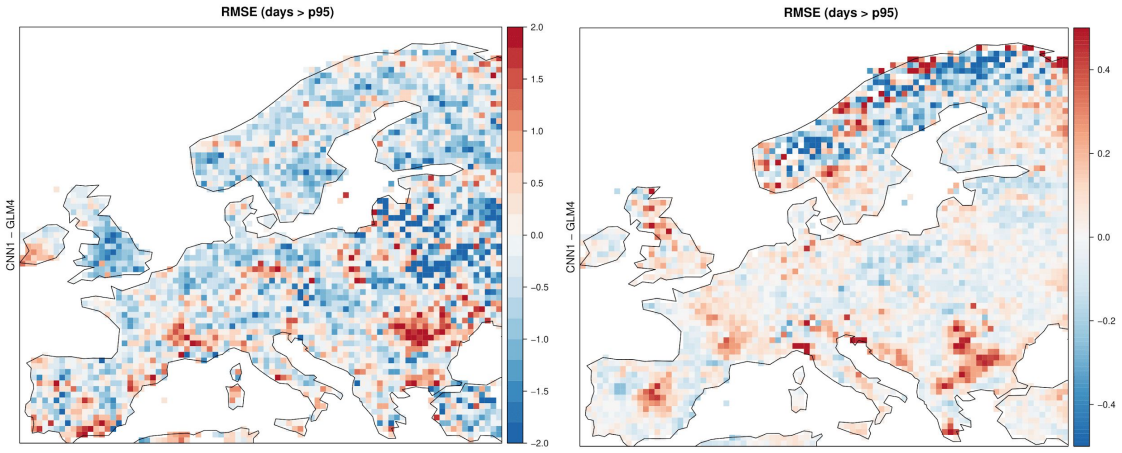
support vector machines (among others). We have introduced a clarification concerning this matter in the new version of the manuscript.

3. A warm validation period is selected as surrogate of possible future climate conditions to investigate the suitability of CNN in climate change studies. It should be better to clearly state how the models produce extremes which are larger than those in the calibration data, and the ability of models to account for changes in the statistics in the future (related to the stationarity assumption)

The choice of a warm test period was done in order to perform a preliminary analysis of the extrapolation capabilities of the statistical models. Figures 4 and 6 in the manuscript suggest that the models intercompared are able to work under unseen conditions during the training phase since the local variability was well reproduced according to the metrics evaluated, especially with the convolutional models (with overall unbiased statistics). We are currently testing the suitability of deep learning approaches to downscale future climate scenarios provided by Global Circulation Models (GCM) and will perform a more detailed analysis related to the stationarity assumption in a future paper. Note that such an analysis is out of the scope of the present paper.

However, to address the referee's concern about the reproducibility of extremes larger than those observed in the training data, we have computed the root mean squared error (RMSE) for those days in the test set for which the observed values were higher than the percentile 95th in the train set, per gridbox. The left (right) column in the figure below shows the RMSE differences between the CNN1 and GLM4 for the case of precipitation (temperature). Red (blue) colors indicate lower RMSE values for the GLM4 (CNN1). For precipitation, CNN1 yields in general lower RMSE values (particularly over southern UK), finding only better results for GLM4 over a limited region in southeast Europe. For temperature, the situation is in general neutral, finding only regional differences in Escandinavia, where CNN1 yields lower RMSE than GLM4. The differences in the RMSE obtained between the methods can be explained

according to whether or not the predictor-predictand link benefits from nonlinearity in the reproduction of extremes.



Differences in the root mean squared error (RMSE) between the CNN1 and GLM4 models (see Table 2 of the manuscript for details in their model setup) for the days in the test set (2003-2008) that have observed values higher than the percentile 95th in the train set (1979-2002) per gridbox. The left (right) column corresponds to precipitation (temperature).

4. Different network architectures of CNN have been evaluated and intercompared in this study. However, the authors should provide more interpretations on the impact of these configuration on model performance. There are a few examples where this is currently done (e.g., lines 215-218, 238-244) but this needs to be done more systematically, and highlighted in the conclusion section.

The differences found in model performance among the deep learning models intercompared in this study depend on the activation function (i.e., linear or nonlinear) and/or the nature of the last hidden layer (i.e., convolutional or dense). For temperature, the predictor-predictand relationship is (quasi)linear and therefore the activation function do not influences the downscaling. In this case, the differences among models are related to the last layer's type of connection: convolutional or dense. For precipitation, the presence of non-linear activation functions benefits the downscaling.

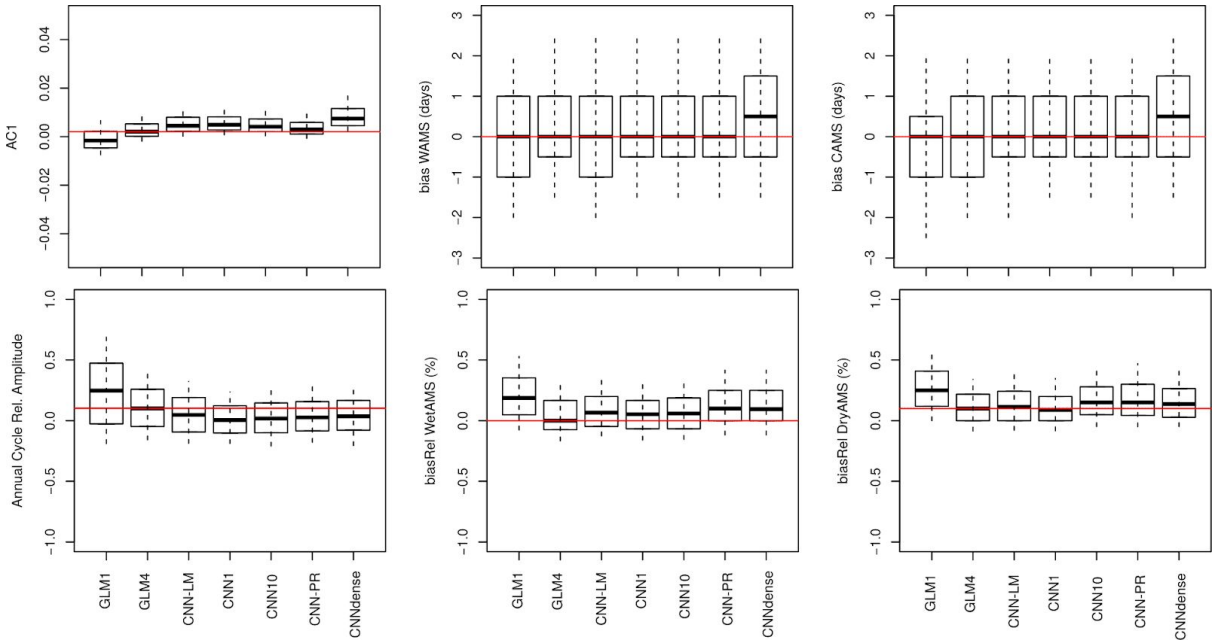
We have highlighted this throughout the new version of the manuscript, with special care in the conclusions section.

5. The skill of the various downscaling methods is assessed mostly on spatial variability. How could the CNN reproduce the temporal variability of the local climate? You may want to validate the ability of CNN to represent dry/wet spells and interannual variation.

Apart from the correlation already shown in Figures 4 and 6 of the original manuscript, we have included the figure below in the new version of the manuscript in order to address this comment. This figure shows three temporal validation metrics for each target variable: temperature and precipitation. For temperature we show the autocorrelation lag-1 and the bias in the length of the longest warm and cold annual spells (first row, from left to right). For precipitation we show the relative amplitude of the annual cycle and the relative bias in the length of the longest dry and wet annual spells (second row, from left to right).

According to these results we can state that no method clearly outperforms the other in terms of reproduction of spells, both for temperature and precipitation. Despite there is some spatial variability (spread of the boxplots) the median results are nearly unbiased in all cases. Only the GLM1 model performs slightly worse for precipitation, which is probably due to the limited amount of predictor information involved in this method. This would indicate that spatial information in the input space is crucial to better reproduce the local variability, which has been already mentioned around Figures 4 and 6 of the original manuscript.

It is worth to mention that any of the methods considered in this work is specifically designed to reproduce advanced temporal aspects such as spells. In the coming future, we plan to explore other battery of methods which explicitly aim to accurately reproduce the observed temporal structure.



Temporal validation metrics computed for temperature and precipitation (top and bottom row, respectively). For temperature, the autocorrelation lag-1 (AC1), and the bias for the length of the longest (bias WAMS) and cold (bias CAMS) annual spells are shown. For precipitation we show the relative amplitude of the annual cycle and the relative bias for the length of the longest wet (biasRel wetAMS) and dry (biasRel DryAMS) annual spells.

MINOR COMMENTS:

1. Line 13: What does 'classic ones' refer to? Need to make them clear.

2. Line 79: 'such'→'such as'

3. Line 111: 'vale' should be 'value'.

4. Figure 2: The label 'bias' is misleading here, since the map shows the differences between the test and train periods based on observations.

We have addressed the minor comments 1, 2, 3 and 4 indicated by the reviewer in the revised manuscript.

5. Figure 4 & 6: The best method is in fact different for each metric, but the same best method (CNN10 for temperature and CNN1 for precipitation) for all metrics is indicated in the figure. How do you choose the best performing method, may be based on one metric?

Figures 4 and 6 show the validation results obtained for temperature and precipitation, respectively. For temperature, the CNN10 is the best method according to the RMSE and the de-seasonalized Pearson correlation while keeping unbiased predictions for the mean, and percentiles 2th and 98th. A similar situation occurs for the CNN1 model for precipitation, for which this method outperforms the others in terms of ROCSS, RMSE and Spearman correlation while getting good results for the rest of metrics. For these reasons we chose the CNN10 and CNN1 models to be the 'best' for temperature and precipitation, respectively.

6. Figure 6: Please explain 'DET'(e) and 'STO'(f).

'DET' refers to deterministic and 'STO' to stochastic. We have clarified it in the new version of the manuscript.

7. Traditional statistical downscaling methods generally require high-resolution observations for model training, thus it is difficult to provide downscaled climate simulations for the regions with little observation data. Is the skill of CNN sensitive to the resolution of observations?

To date we have only used deep learning to downscale to resolutions of 0.5º and, despite the sensitivity of downscaling to the observational reference considered is a relevant topic of study, it is out of the scope of this paper. However, we hypothesize that the sensitivity of the downscaling to the predictand's resolution is mainly related to the explicability of the local scale by the predictor's domain rather than by the downscaling method itself (e.g., convective precipitation is not explicable by large-scale predictors and therefore the ability to establish a robust link between both is independent of the statistical method of choice). The benefits of convolutional approaches, such as the ability to treat high-dimensional domains without previous feature selection techniques and the ability to

extract non-linear patterns from data, are intrinsic to the multisite-convolutional nature and therefore the latter skills are expected to be preserved indistinctly of the predictand's resolution. In fact, downscaling to higher resolutions may require a higher degree of nonlinearity and therefore, the skill of deep learning could be even increased in comparison with classical approaches.

In the case of regions with scarce observations, computer vision applications have benefited from a concept called "transfer learning". The idea behind transfer learning is that hidden features learned in a particular task *A* are useful in a similar task *B* and therefore, the trained network *A* (or the first hidden layers) can be used to predict task *B*. In the case of downscaling, though this has not yet been tested to our knowledge, a net trained over a well observed region (e.g., Europe) could be transferable as a pretrained-net over areas with less observations available (e.g., Arctic). Though there are still questions to be answered in this topic such as whether the hidden features learned to downscale temperature over Europe (even the most simple ones located in the first hidden layers) would be helpful to downscale in regions with scarce observations which can present their own climatic particularities.