1. The authors should say something about the computational effort of the proposed methods, saying if there is any trade-off between performances (RMSE, correlation, etc.) and complexity/computation time. I expect that a linear model should run much faster than a CNN, can the authors say something about this?

We have used the set of packages *climate4R* (https://github.com/SantanderMetGroup/climate4R) for the linear models and the *downscaleR.keras* package (https://github.com/SantanderMetGroup/downscaleR.keras), which integrates keras into climate4R, for the deep models. We have isolated the code needed to train and predict in the test set for both generalized linear models (GLM) and deep models and retained the computation times (see the table below). It must be noted that for precipitation there are two generalized linear models to train (a binomial logistic and a gamma logarithmic for the occurrence and amount of rain, respectively) and therefore, the time included in the table for GLM1 and GLM4 is the sum of these two individual GLMs. Differently, in deep learning models the occurrence and amount of rain are trained simultaneously. In this case, the speed of training is very sensitive to some parameters such as the learning rate (learning rate equal to 0.0001) and the early-stopping criteria (patience with 30 epochs) which mainly drive the number of epochs or iterations needed to train the model. Taking into account all these considerations we observe little difference between the computational times needed to train the linear models (GLM1 and GLM4) and the deep ones (CNN1; the rest of deep configurations yield similar computing times). Therefore, the computational effort required to train and run deep models is not a strong limitation.

	GLM1	**GLM4**	**CNN1**
precipitation	47	80	74
temperature	22	28	62

Computational times (in minutes) needed to train the GLM1, GLM4 and CNN1 methods (see Table 2 of the manuscript for information about the configuration of the models) for precipitation and temperature.

Based on this referee's comment we have included an annex in the new version of the manuscript which shows the computational times required by the different methods.

2. Possibly related to the point 1. probably: the authors use different CNN setups but then they analyse only the best one (CNN1), can they say something about the others? Why they do not work well? Why they were supposed to work well? Why they are considered in the paper?

Unlike other disciplines where deep learning is well established, the "black box" character of neural networks is a major concern in the earth sciences community. In contrast to other deep learning and downscaling studies where complex computer vision topologies are adopted without a proper justification, in this study we propose an intercomparison among deep models of increasing levels of complexity in order to shed light on the role of the different elements involved in this kind of approaches for downscaling. For instance, we consider a convolutional model with linear activation functions (CNN-LM in the manuscript) and its equivalent with nonlinear activation functions (CNN1). This allows to analyze the influence of nonlinearity in the downscaling. Our results show that the introduction of nonlinearities in the model is relevant for precipitation but not for temperature. CNNdense and CNN-PR (see Table 2 of the manuscript for the details) were included in the study since this type of networks are often used in computer vision applications and we wanted to test their potential suitability for statistical downscaling purposes. Whereas CNNdense results from the idea of mixing the spatial patterns learn by the convolutions in the last hidden layers, CNN-RPR is based on the idea that more filter maps are needed as we go further in the net, given the increase of nonlinearity. Both methods obtain similar results to GLM4 and are clearly outperformed by only-convolutional topologies due to the spatial dependence of the predictand's output neurons in the last hidden layer (i.e., the prediction over a particular site its dependent on the atmospheric situation surrounding that

area and thus mixing the spatial patterns in dense layers damages the downscaling).

A description of the deep models proposed and why they are considered in the current paper (lines 157-170) can be found in Section 3 of the manuscript. All deep learning models have been intercompared in Figures 4 and 6 in terms of all the proposed metrics for temperature (RMSE, Pearson correlation, bias of the mean, percentile 2 and percentile 98 and the ratio of standard deviations) and for precipitation (ROCSS, RMSE, Spearman correlation, bias for the mean and percentile 98). The spatial maps displayed in Figures 5 and 7 were only shown for the generalized linear models (GLM1 and GLM4) and for the best method (CNN1 and CNN10 for precipitation and temperature, respectively) as no additional conclusions than those inferred from Figures 4 and 6 appeared when considering the other deep models.

3. For the precipitation the authors use a probabilistic score (ROCSS) in addition to the common ones (RMSE, Correlation, etc), it's not clear how the output of a linear model or a CNN could be considered a probabilistic forecast. They should clarify this point.

Note that the ROCSS in only used for the binary (0/1) event *occurrence of precipitation*. In GLMs, there is a first GLM with binomial error distribution and logit link function whose outputs can be directly understood as probability of rain for a given day at a given gridbox.
Likewise, CNNs minimize the negative-log-likelihood of a Bernouilli-Gamma distribution (see Figure 3 and lines 170-178) providing therefore an estimation of the parameters *p*, *alpha* (shape parameter of a gamma distribution) and *beta* (scale parameter of a gamma distribution), simultaneously. *p* is the probability of rain for a given day at given gridbox.
Therefore, both GLM and CNN models provide the needed information to compute the ROCSS.

More information can be found in the paper notebook (2019_deepDownscaling_GMD.pdf), where there is a step-to-step explanation of the results presented in the paper.

4. Can the authors comment (or provide reference) on how they decided the best configuration for the CNN? Number of layers, etc. This could be beneficial especially considered that the journal is for a community that, as you say, does not really trust deep learning models.

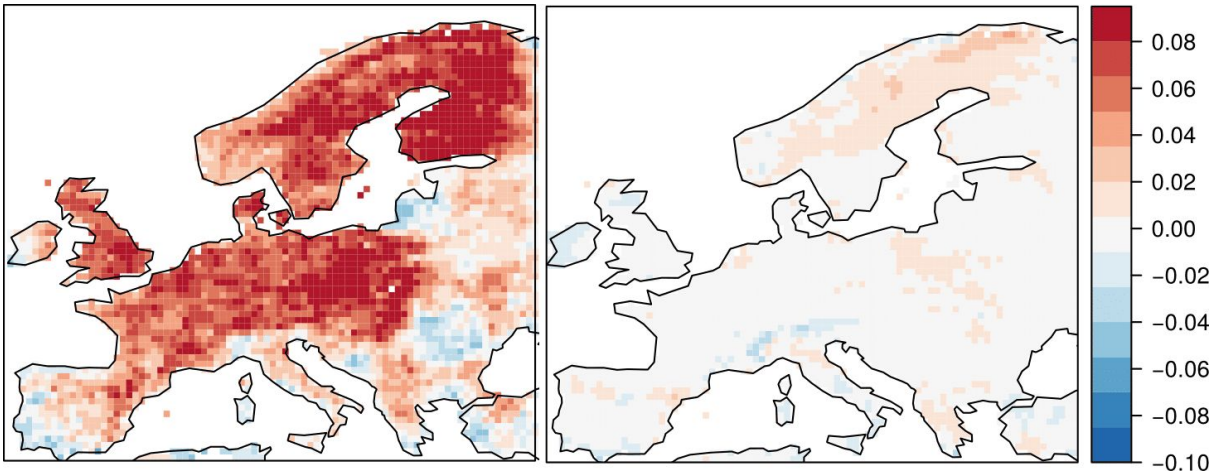
The number of layers depends mainly on two aspects: the degree of nonlinearity you want to achieve in your model (the deeper the more nonlinear) and the number of parameters involved in your model. Unlike computer vision applications in which there are usually more than 50.000 images available for the training phase, here we only had 24 years of daily data. As a consequence, the depth of our networks is limited. Though it was not discussed in the manuscript we tried different topologies that varied mainly in the number of layers (up to 6) and in the kernel size (we used for the paper 3x3 kernels but also tried 5x5 and 7x7 sizes). After this trial and error procedure we ended up with the optimum of 3 convolutional layers and a 3x3 kernel size. Therefore, additional layers seem to not benefit the model due to an overparameterization when no more nonlinearity is actually needed. Likewise, the final choice of kernel's size being equal to 3 is related to the fact that most relevant phenomena for downscaling at the resolution considered in this work occurs in a surrounding domain of 3x3, with bigger domains just adding unnecessary degrees of freedom to the model.

5. Regarding the comment about deep learning and distrust in climate community, I have the impression that the problem is not just about the extrapolation capabilities, but in general about the impossibility to really know how a black-box model operates. The extrapolation is only a part of it. You can not really assess the capability to "extrapolate" for a complex model like a CNN because any assessment would be 1. configuration specific and 2. data specific. Then I think the problem is a conceptual one: the difficulty in generalising the behaviour of a very complex and highly-nonlinear model. (This point is just a personal comment, I think that this paper is not the right place to for this kind of discussion however I have really appreciated that comment)

Thanks for your comment, we find interesting your discussion about the “black box” nature of neural networks. It is true that the extrapolation capability is only a part of it and further efforts have to be done with regards to other issues such as as the quantification of uncertainty in the predictions by (deep)bayesian approaches or the sensitivity to the choice of predictors.

6. Can the authors provide a map with the difference between metrics (RMSE or correlation) between GLM4 and CNN1? Can they say something about the areas where CNN/GLM outperforms the other method?

The figure below shows the differences in de-seasonalized correlation found between CNN1 and GLM4 for precipitation (left column) and temperature (right column). Red (blue) colors indicate that CNN1 yields higher (lower) correlations than GLM4.
For precipitation, better results are found for CNN1 over most of Europe, especially in Escandinavia, the British isles and central Europe. Eastern Europe and the Mediterranean differ slightly among models. This may be due to the fact that the “true” predictor-predictand link is quasi-linear in those areas and CNN has little added value apart from the automatic treatment of the input space. This conclusion is also applicable to the results found for temperature, for which the “true” linear existing relationship is again (quasi-linear and CNN1 and GLM4 do not show significant differences in terms of correlation.



Differences in de-seasonalized correlation for precipitation (left) and temperature (right) found for CNN1 and GLM4 (the latter is taken as reference).

7. The DOI at line 72 does not work. There is a typo in the first panel of Figure 1, in the caption title.

We have updated the version and the correct DOI will appear in the new version of the manuscript.