Manuscript:

"Deep learning models for generation of precipitation maps based on Numerical Weather Prediction"

Response to editors and reviewers

Dear Editors and Reviewers,

We thank and appreciate all the comments made to our paper, as well as your positive observations about the relevance and merit of the manuscript. We are sure these will help improve the quality and readability of the paper. In the following, we provide an answer to each comment in blue, followed by an indication of the changes made in the manuscript.

The authors,

Chief editor comments

Dear authors,

We have detected some missing material in your manuscript. Since your work is an application of deep learning techniques, we need that you include with your manuscript the input and output files used and obtained. You state that you are using as input data output files from COSMO. We do not need the complete output files from COSMO but the fields and variables you use. This is important information to assure minimum replicability of your results, as they depend critically on the input and much more than works using other models.

Therefore, please, publish the data in one of the appropriate repositories (b2share is ok), and reply to this comment with the relevant information (link and DOI) as soon as possible, as it should be available for the Discussions stage.

Also, in this way, you must include in a potential reviewed version of your manuscript the modified 'Code and Data Availability' section, with the DOI of the data.

Juan A. Añel
Geosci. Model Dev. Exec. Editor

We have uploaded the requested datasets to the directory https://doi.org/10.5281/zenodo.7244319 and an identic copy of the code on the GitHub repository can be downloaded in https://doi.org/10.5281/zenodo.7535434. This information was added to the manuscript in the "Code and Data Availability" section.

Editor comments

In this study, the authors attempt to enhance low-resolution precipitation forecast maps from the NWP model by using deep learning models. Several deep learning models used in this study are examined, which include UNet, 2 deconvolution networks, CGANs, and a baseline. Their results demonstrate that direct mapping between physical simulations and precipitation maps can be achieved by DL models. However, the accuracy of predicting precipitation maps using their ML methods is still a challenge. Overall, both the merit and the approach used in this study are interesting and worth consideration for publication on EGUsphere. I have only a few minor questions to help readers better follow the significance and methodology presented in this study.

1. It is mentioned in this study that 143 forecast variables are used as input for the deep learning models in this study. However, there is nowhere in the text mentioning what are these variables and why are they relevant to precipitation augmentation. I am wondering what are the possible 143 different variables that a NWP model can produce, and how these variables are handled in your DL approach. Are they treated as different channels of input? If not, how are they combined and/or fed in your DL designs? It would be useful for readers to know more details about what these variables are and how are they used as input for your DL models.

The 143 variables correspond to the ensemble statistics (mean and standard deviation) of the 20 ensemble members of the COSMO-DE-EPS, which represent simulations for several atmospheric variables (wind speed, temperature, pressure, etc.) and soil and surface variables (water vapor on the surface, snow amount, etc.). As correctly assumed, we provide the different variables as input channels for the DL model (we tried to illustrate this in Figure 2). The DL models combined the input channels in a non-linear fashion using the deconvolutional kernels to generate high-definition precipitation maps.

We rewrote the first paragraph of the data section to include information about the input variables in lines 85-88, and details about the combination performed by the algorithms were included in lines 124-125.

2. Precipitation is generally a subtle variable, which is an end product of many processes and scales in a numerical model. While the overall objective of this work is to enhance the model precipitation output by using ML, the characteristics of different types of precipitation such as stratiform or convective precipitation are very different. I am not sure if the ML models can help distinguish these different types of precipitation, which is in fact related to my comment # 1 above on the use of 143 forecast variables as input. There is no discussion of how many of these variables are essential for different types of precipitation. One could of course combine all possible input and see what potential outcome from an ML model could be. However, more input channels do not generally lead to a better outcome, since some bad channels could degrade the ML performance. Any discussion on the relative importance of different input variables for different types of precipitation would be helpful here.

This is an important observation. Our approach was to include sufficient input information about the atmospheric and soil state (143 channels), together with a sufficiently complex model, and let the model automatically discover the relevant patterns in the input data that improve the precipitation forecast. This means that the complex DL models learn the relevant interactions between the input information for each type of precipitation without an explicit classification from our site for the types of rain. Given the improvement in the performance obtained by our models, we could expect that the

DL algorithms use the proper and filter out the unimportant information in each case, however, there is no proof for that more than the good performance on the test set.

We have added two paragraphs to the discussion concerning this limitation in lines 379-391.

3. As a "model description paper", the manuscript is expected to be detailed and accessible for a wide range of geophysical communities as described here https://www.geoscientific-model-development.net/about/manuscript_types.html#item1. However, the current methodology section (section 3) is too brief for readers to follow and appreciate your ML model settings and approach. Please provide additional information as instructed in the link above to meet the standard guideline of EGUsphere.

We add a new section to the Methodology where we concentrate on all the implementation details previously distributed throughout the paper. We add to this section a detailed description of the implementation together with all the technical and numerical details needed for the reproducibility of our findings (see lines 184-224). Additionally, we add a version number to the original models presented by the paper (Deconv1L4Rain, Deconv3L4Rain, and CGAN4Rain), following the indications of the first point of the link.

Referee 1 comments

1. Introduction section

(1) there is confusion about scientific question. Did it compare different model algorithms or generate precipitation map or both?

The main goal of the paper is to generate high-resolution precipitation maps, and to achieve that, we compared the algorithms in their performance to obtain information about how to solve best this specific problem. Our results show that the CGANs generated the highest-quality precipitation maps. We made this explicit in line 62.

(2) What is the novelty of this study? Just a complete set of variables?

The novelty of this study is to use the complete set of variables from the atmospheric simulation and the combination of two steps into one: downscaling the precipitation forecast and correcting its bias with a single algorithm. This was pointed out in the text in line 57.

(3) There is a lack of some references in this section, such as line 22 "Among the most successful methods are the Numerical Weather Predictions models (NWP), which consist of systems of equations that simulate the dynamics of the atmosphere and provide highly accurate weather forecasts over long periods." Line 27 "However, NWP models still preserve some limitations, the most important being the large number of computational resource needed to generate forecasts" and so on. Please double check the whole text.

In reality, all the information had one common reference (Serifi et al., 2021). The paragraph was rewritten to show this (see lines 27-29).

1. Data section

(1) I think it is not well-structured. There is no information about study area, and some information lacks of reference. Thus, I do not know is it your result or others, such as “The south-western part is characterized by low precipitation amounts between 500 and 700mm/yr due to lee effects of the Eifel mountain range”.

We rewrote the mentioned paragraph and integrated additional information in the text, as well as in the respective figure. We also include additional references. See lines 71-80.

(2) Please give more descriptions about COSMO-DE-EPS forecast, providing more information or reliability for readers.

We rewrote the paragraph to add additional information about the COSMO-DE-EPS. See lines 86-90.

(3) Please give full name of “RW, YW, RQ” when they are first used in this study.

These abbreviations are official product names of the RADKLIM data, and the original references do not clarify what the letters represent. Adding them to the paper is unfortunately out of our hands.

1. Methodology section

Please give more descriptions about five deep learning models, such as:

(1) Why choose these five models in this study?

(2) How to set up these models in this study? For example, how to choose variables? How to calibrate and validate model parameters? How to deal with the correlation of independent variables?

(3) The setup is done by this work or referred from other study?

We add an additional paragraph to the data section answering these questions. See lines 144-153.

1. Result section

In this section, the key information is from Table 2 and Figure 3. But it confused me that COSMO-DE-EPS data in Table 2 is original data or after correction? Maybe I missed some information. If it had been corrected by observation, it is not surprised that it has good performance.

Table 2 presents the scores of the COSMO-DE-EPS original data. It is actually a good original performance. But the deep learning models that we trained based on COSMO-DE-EPS obtained a significant improvement in their fidelity in comparison with the original COSMO-DE-EPS.

1. Conclusions section

In this section, authors presented a summary of results. Maybe authors can discuss some uncertainties about five models, such as the parameters and the influences of model uncertainty on precipitation generation results.

An additional paragraph with considerations about the limitations of the models was added to the discussion. See lines 379-386.

Referee 2 comments

This is an interesting and useful study examining the comparative ability of several neural network (NN) architectures to improve short-term/fine-scale precipitation forecasts from NWP models. The paper demonstrates measurable improvements relative to the baseline while also highlighting clear limitations in the ability of any of the architectures to correctly predict more intense rainfall. I believe this paper is suitable for publication with the following suggested minor revisions.

Line 35: Please expand the abbreviation 'GRU'

We substitute the abbreviation 'GRU' with 'recurrent' which is more general and still captures the central concept of the model. See line 35.

Line 81–84: It seems unlikely to me that all 143 forecast variables contribute significantly to the rain prediction. I wonder whether training speed and robustness — including reducing the risk of overfitting — could be significantly improved by substantially reducing the variable set. I don't expect the authors to perform new runs for this paper, but some comments on the question of "how many variables is too many?" would be welcome.

The question about the number of variables is an important one. We consider the strength of our work to be the mixing of multiple variables, including additional information about the meteorological state, to generate the rain maps. However, as pointed out, this involves more computational resources during training. We address this comment in the discussion in lines 379-391.

Section 2.3: It would be helpful to know what fraction of the roughly 20,000 observations had any rain at all. An overall histogram or CDF of rain rate for the entire observational data set could also be useful. If the fraction of rain is low, then I would imagine that the NNs are effectively being trained to not predict rain by default, which might help explain the low biases? Would it be desirable to bias the training sample toward raining cases to reduce this tendency?

This was a critical faced difficulty. During training, if the right hyperparameters were not selected, the model got stuck in a "local minimum" of predicting zero rain in all cases (which assures an accuracy of around 85%). Fortunately, we overcame this local minimum by experimentation and finding the correct hyperparameters for the problem. None of the reported models was stuck in this local minimum of predicting only zeros. Even if the mentioned are well-known resampling techniques, training the data oversampling rainy data or with an artificial distribution of precipitation could induce a new type of bias. Therefore, we decided to train the algorithms with naturally distributed data (which also makes our approach more robust). We add Figure 3, which illustrates the distribution of rain from the datasets.

Line 135: If I'm reading this correctly, all models were trained for a fixed number of epochs without early stopping and without specific tests for convergence. Is this correct? Please consider clarifying the training procedure with a few additional details.

Yes, all models were trained for a fixed number of epochs without using early stopping. However, the loss during training was plotted at the end of training to check for convergence. Additional details were added in line 218-219.

Line 137: Was an optimal learning rate determined experimentally? Is it to be expected that the same learning rate is optimal for several very different NN architectures?

In this case, yes. All hyperparameters were inspired by literature but tested with extensive experimentation. Additionally, using equivalent hyperparameters allowed for direct comparison between the different models.

Section 3.3.3: Heading should be "Equitable Threat Score"

Fixed! Thanks. See line 250.

Table 2: Finding the highest median score in each row, I see that Deconv3L did best for 0.2 mm/hr, Deconv1L did best for 0.5 mm/hr, and U-Net did best for 1 mm/hr. And all three did better than CGAN for these rain rate thresholds. Yet CGAN did best in MAE and LEPS scores. The lack of a consistent pattern of any one method standing out by most or all metrics seems to me to be a significant finding, but I don't think this comes through clearly as clearly as it could in the discussion or in the conclusions.

We based the overall evaluation of the models on the MAE and LEPS skill scores. The better performance of the different models on the beforementioned thresholds shows the mid-range precipitation bias of the U-Net and the spatial smoothing of the deconvolution. Given the strong gamma-like-shaped distribution of precipitation, a tendency towards mid-range precipitations generates a loss in the general performance (as we measure it with MAE and LEPS). The CGAN model shows a significant performance through the majority of the thresholds (even if, as mentioned, it is not the best for some specific thresholds), showing, in our opinion, a generally superior performance compared to the rest of the models, offering guidelines for future research about which type of algorithm to use.

Line 236: Should "overcast" be "overforecast"?

Corrected. See line 283.

Line 272: "Considering the excellent performance shown by the Deconv3L model …" I suggest "superior" (in the sense of "better by some metrics than the other models") rather than "excellent", as the latter word implies that there's only limited room for improvement. The MAE score of 0.5 strikes me as good in relative terms but not entirely satisfying in absolute terms.

We changed "excellent" to "superior" as suggested. See line 321.

Lines 334-335: To “future research”, I suggest including an evaluation of which of the 143 forecast variables are actually relevant to the models’ performance—e.g., by withholding one at time and seeing how performance changes. As mentioned above, I suspect that excluding irrelevant features from the data set—and thereby improving the signal-to-noise ratio at the inputs—might improve the robustness of the training that is possible with a limited data set.

This comment coincides with an observation made by the editor. We included an additional paragraph in the discussion. See lines 379-391