

Towards replacing precipitation ensemble predictions systems using machine learning

Rüdiger Brecht[†] and Alex Bihlo[‡]

[†]*Department of Mathematics, University of Hamburg, Hamburg, Germany*

[‡]*Department of Mathematics and Statistics, Memorial University of Newfoundland, St. John's (NL) A1C 5S7, Canada*

E-mail: ruediger.brecht@uni-hamburg.de, abihlo@mun.ca

Precipitation forecasts are less accurate compared to other meteorological fields because several key processes affecting precipitation distribution and intensity occur below the resolved scale of global weather prediction models. This requires to use higher resolution simulations. To generate an uncertainty prediction associated with the forecast, ensembles of simulations are run simultaneously. However, the computational cost is a limiting factor here. Thus, instead of generating an ensemble system from simulations there is a trend of using neural networks. Unfortunately the data for high resolution ensemble runs is not available. We propose a new approach to generating ensemble weather predictions for high-resolution precipitation without requiring high-resolution training data. The method uses generative adversarial networks to learn the complex patterns of precipitation and produce diverse and realistic precipitation fields, allowing to generate realistic precipitation ensemble members using only the available control forecast. We demonstrate the feasibility of generating realistic precipitation ensemble members on unseen higher resolutions. We use evaluation metrics such as RMSE, CRPS, rank histogram and ROC curves to demonstrate that our generated ensemble is almost identical to the ECMWF IFS ensemble.

Key points:

- We propose a novel method for generating precipitation ensemble members from deterministic weather forecasts.
- Prediction is done using a generative adversarial network in an image-to-image style.
- The neural networks is trained on low resolution data but can be applied on unseen higher resolution data.

1 Introduction

Precipitation forecasting is an essential aspect of weather prediction, with significant implications for various sectors, such as agriculture, transportation, water management, and disaster preparedness [21, 28]. In recent decades, the quality of numerical weather prediction has significantly improved, with meteorological centers utilizing numerical models and reanalyses with grid spacing ranging from 10 to 80 km and updated in near real-time [5]. However, while these grid spacing are capable of resolving large-scale weather patterns, they are insufficient for accurately representing precipitation in regions with subgrid-scale orographic variations [15]. Thus, precipitation intensity can exhibit significant variations over short distances, with scales of 1 km or less, which is far finer than the typical resolution of global weather models. It is imperative

to enhance the resolution of precipitation forecasts to accurately evaluate potential impacts, particularly for scenarios involving extreme rainfall.

Moreover, ensemble weather prediction aims to quantify the different sources of uncertainty in numerical weather prediction models, cf. [18]. The most significant sources of uncertainty are the initial conditions and errors in the numerical model formulation. To address these uncertainties, an ensemble of perturbed forecasts is generated, in addition to the single deterministic weather forecast, with the overall divergence, or spread, of the ensemble ideally providing a measure of the uncertainty in the deterministic prediction. However, the main constraint in generating the ensemble is still computational, as each ensemble run requires significant computational resources, thereby limiting the total number of ensembles that can be computed on an operational basis, to typically less than 100.

This article presents a generative deep learning approach to generating ensemble weather forecasts for high-resolution precipitation forecasts, without being trained on high-resolution data. We use generative adversarial networks (GANs) to learn the complex spatio-temporal patterns of precipitation and generate diverse and realistic precipitation fields.

Thus, the main contributions of this paper are:

- Using machine learning to generate realistic precipitation ensemble members from using only the available control forecast;
- Showing that it is possible to generating realistic precipitation ensemble members on unseen (higher) resolutions using lower resolution training data.

The remainder of this paper is organized as follows. In Section 2, we provide a brief overview of related work on deep learning in meteorology and ensemble forecasting for precipitation. Section 3 is devoted to the description of our proposed method using a generative deep learning approach. In Section 4 we present the results and verification of our approach. Finally, in Section 5, we present a summary of the work carried out and discuss future directions for research in this area.

2 Related work

With the advent of modern deep learning following the seminal contribution [17], neural networks have become a topic of immense research interest in meteorology. Neural networks have been used for precipitation nowcasting [23], data-driven weather forecasting [1, 20, 25], parameterization [7], as new numerical methods [2], and data assimilation [6]. Deep learning based models have also been used to explore a range of data-driven ensemble prediction methods. In [1], a conditional generative adversarial network was used to predict various meteorological quantities, with the ensemble prediction model being generated using Monte-Carlo dropout. In [22], several methods for generating ensemble predictions using deep learning were investigated, which included random perturbations, retraining of neural networks, Monte-Carlo dropout, and ensembles based on singular value decomposition. The work [9] explored the use of deep convolutional neural networks to reduce the number of ensemble members while preserving the correct ensemble spread, while [10] proposed post-processing computed ensemble members using neural networks. Ensemble forecasting based on convolutional neural networks on a cubed sphere for sub-seasonal forecasting was suggested in [26]. The ensemble generation by stacking sub-neural networks trained on different subsets of predictive meteorological variables was put forth in [4]. A framework based on physically constrained generative adversarial networks to simultaneously

improve local distributions and spatial structures of ensemble predictions was proposed in [14]. The possibility of directly generating ensemble weather forecasts from the deterministic weather prediction only was shown in [3], who successfully used it to obtain ensemble weather predictions for the 500 hPa geopotential height. Crucially, this approach only generated the ensemble spread, not individual ensemble members. It is this approach that will be extended here to a full precipitation ensemble prediction system, whereby the control forecast is used to generate different ensemble members, rather than just the spread derived from these members.

Finally, also closely related to our work here, in [12] different generative neural network architectures are compared to stochastically down-scale a precipitation forecast. This way a high resolution ensemble forecast can be generated. However, this method relies on high resolution radar data samples, which are not available for all spatial locations where ensemble forecast would be needed.

3 Methods

In this section we describe the neural network architectures and the data being used to train these networks.

3.1 Data

We use the operational ECMWF IFS precipitation data. We downloaded the data on a regular longitude–latitude grid over northern Europe, with latitude $\lambda \in [45, 73.5]$ and longitude $\varphi \in [-27, 33]$ with a grid resolution of 288×513 grid points. We retrieved the ensemble data initialized at 00 UTC, for the next 72 hours at a temporal resolution of 6 hours corresponding to 12 time steps per ensemble run, from the years 2020–2022. We use the years 2020 and 2021 for training and 2022 for testing. For training we randomly split the data into 95% for training and 5% for validation.

We note here that for deep neural networks is a rather modest training dataset. However, based on the previous work in [1, 3, 12], where training datasets of comparative size were used, we also found for the present study that this already allowed us to train highly performing neural network models.

3.2 Neural network architecture

Our method is such that each model accepts as input the deterministic precipitation forecast for the next 72 hours, and predicts a single ensemble member for the same time period. We then train a total of 10 such models, which collectively form our ensemble prediction system. We have chosen to generate 10 ensemble members rather than 50 members, as is done operationally at the ECMWF, since this provided a good compromise between training effort and statistical variability of the ensemble.

Alternatively, one could train a single model with a total of 10 output heads, but we found experimentally that the former method gives rise to a more diverse ensemble. While our method incurs a substantially higher training time, training only has to be done once and can be trivially parallelized. Our method also allows us to add further ensemble members without having to retrain the models for the existing ensemble members.

We interpret the precipitation forecast as a set of 12 image snapshots and note that the precipitation at a given time is the cumulative precipitation until that time. We consider a model

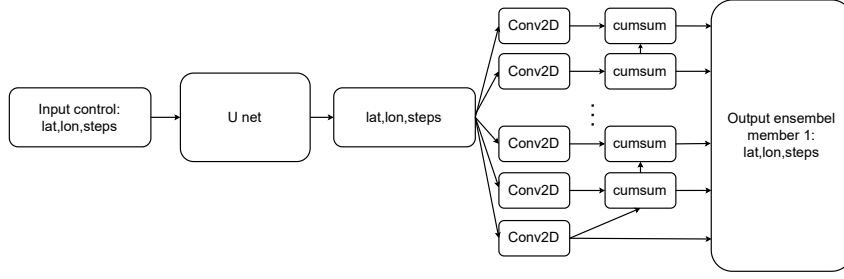


Figure 1. The neural network architecture of the generator network. For the architecture of the U-net, see Table 1. The output of the U-net is fed into 12 parallel convolutions, one for each time step, with a single filter of size 3 and a ReLU activation. To guarantee that the cumulative rain is increasing over each time step, the output at each time step is formed by adding the respective convolutional output to the previous cumulative output.

that is based on the classical `pix2pix` architecture adapted from [16], where it was used for a variety of image-to-image translation problems. This model was selected as it has successfully been used in the related works [1, 3]. It constitutes a conditional generative adversarial network, which consists of a generator and a discriminator network, which are trained in tandem through a two-player minimax game, see [8] for more information.

The neural network generator architecture is visualized in Fig.1. The discriminator architecture is the same as those used in [1, 3], with 2D convolutions instead of 3D convolutions and is thus not shown here. The details¹ of the architectures of both the generator and discriminator can be found in tables 1 and 2.

We train for each of the 10 ensemble members a neural network on both low resolution input (control forecast) and output (ensemble) data to predict the respective ensemble member given only the control forecast as input. The evaluation of the trained network is then done on the full high resolution input and output data.

Our reasoning behind using degraded data is to mimic the unavailability of high-resolution ensemble runs. We also note that as operational forecasting centers move towards higher resolution models, the respective ensemble predictions become computationally more expensive as well. If an ensemble prediction system pre-trained on lower resolution data still performs well on higher resolution data, this can help ease the computational and environmental burden of such systems.

Practically, since we use a convolutional neural network for the generator, which does not require a specification of the spatial input data dimensions, the trained neural network can be applied to arbitrary resolution data. The only constraint here is that the input and output data have to be integer multiples of the original training data. In the following, we will experiment with training our model on data that uses only every second and every third grid point, respectively, while being evaluated on the full high resolution testing dataset.

3.3 Training

We train three sets of neural networks, in the following referred to as *nn1*, *nn2* and *nn3*, respectively, one for the full resolution, one for degraded data where we take every second grid point and one where we take every third grid point. We implement the models in TensorFlow 2.8 and trained them simultaneously on a machine with a dual NVIDIA RTX 8000 GPU. We use the Adam optimizer with a learning rate of $2 \cdot 10^{-4}$ using a batch size of 1. We trained the

¹The code can be found on [Github](#) upon request.

Table 1. Implementation details of the U-net neural network. The first three deconvolution layers are concatenated with the output of the convolution layers to form the skip connections.

U-net	no. of filter	kernel size	strides	activation
2D convolution	64	4	2	LeakyReLU
batch normalization				
2D convolution	128	4	2	LeakyReLU
batch normalization				
2D convolution	256	4	2	LeakyReLU
batch normalization				
2D deconvolution	256	4	2	ReLU
batch normalization				
2D deconvolution, 50% dropout	128	4	2	ReLU
batch normalization				
2D deconvolution, 50% dropout	64	4	2	ReLU
batch normalization				
2D deconvolution	12	4	2	ReLU

Table 2. Implementation details of the discriminator neural network.

Discriminator	no. of filter	kernel size	strides	activation
2D convolution	64	4	2	LeakyReLU
2D convolution	128	4	2	LeakyReLU
2D convolution	256	4	2	LeakyReLU
2D convolution	512	4	1	LeakyReLU
batch normalization				
2D convolution	1	4	1	Linear

models for 30 epochs using early stopping, saving the model after each epoch, and chose the one with the lowest validation error. One training epoch for all models took about 20 minutes.

3.4 Ensemble verification methods

We evaluate how well our model can reproduce the IFS ensemble by evaluating how similar the respective ensemble distributions are. Specifically, we choose as error measures:

- The root mean squared error (RMSE),
- The continuous ranked probability score (CRPS) [13],
- The Receiver Operating Characteristic (ROC) curves, and
- The rank histogram.

Below we give the definition of the RMSE and the CRPS,

$$\text{RMSE}(x, y) = \sqrt{(x - y)^2},$$

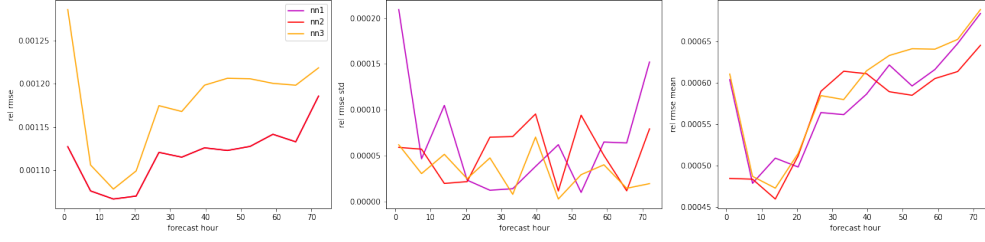


Figure 2. Forecast verification RMSE over the test dataset as a function of the forecast step.

$$\text{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(x) - \mathbb{1}(x - y))^2 dx.$$

Here, the overbar denotes spatial averaging and F is the cumulative distribution function of the forecast density function, and $\mathbb{1}$ is the Heavyside step function. The smaller the RMSE and CRPS the better the result is.

The ROC curve evaluates the effectiveness of a binary classifier by plotting the true positive rate (sensitivity) against the false positive rate ($1 - \text{specificity}$) over a range of probability thresholds [27]. In order to create ROC curves for a specific precipitation intensity, we generate an ensemble of predictions. For each pixel, we determine the percentage of ensemble members that predicted rainfall above the specified intensity, which we interpret as the probability of the event occurring. Each point on the ROC curve represents the performance of the system at a particular probability threshold used to differentiate positive from negative predictions. The ROC curve summarizes the system’s performance across all probability thresholds, ranging from 0 to 1.

The rank histogram [11] is generated by repeatedly tallying the rank of the verification (usually an observation) relative to values from an ensemble sorted from lowest to highest. A well-calibrated ensemble results in an uniform histogram. Here, however, we are only interested in verifying that the generated ensemble has a similar histogram compared to the IFS forecast ensemble, as this is the data our models were trained on. Improving the bias in the ensemble prediction system itself is beyond the scope of this work.

4 Results

In this section we demonstrate that it is possible to learn the ensemble members on a low resolution, enabling to predict the associated ensemble members on a higher resolution. For this we evaluate the trained neural networks on the unseen high-resolution data of 2022.

4.1 RMSE

We compute the RMSE between the IFS ensemble and the ensemble generated for each of the 12 lead times and average the RMSE over the year. We also compute the RMSE between the ensemble mean and standard deviation. In Fig. 2 we can observe that the error between the IFS ensemble and generated ensemble does not differ much between the low resolution (where it is trained on) and the higher resolution.

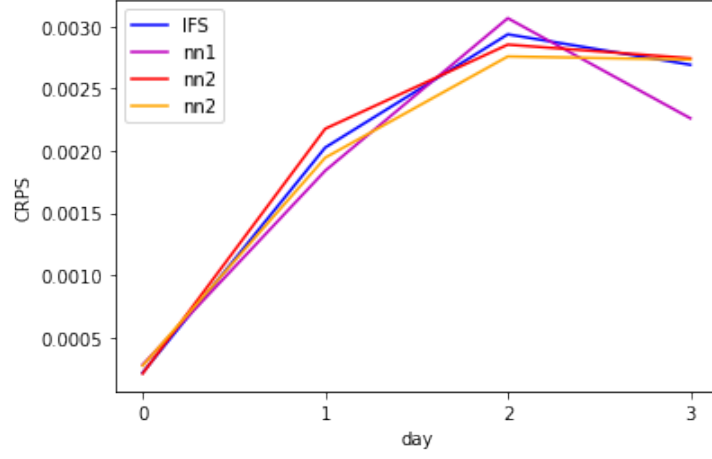


Figure 3. CRPS scores for the generated and IFS ensembles after 0, 1, 2 and 3 days with the control run step 0 at those days as a reference.

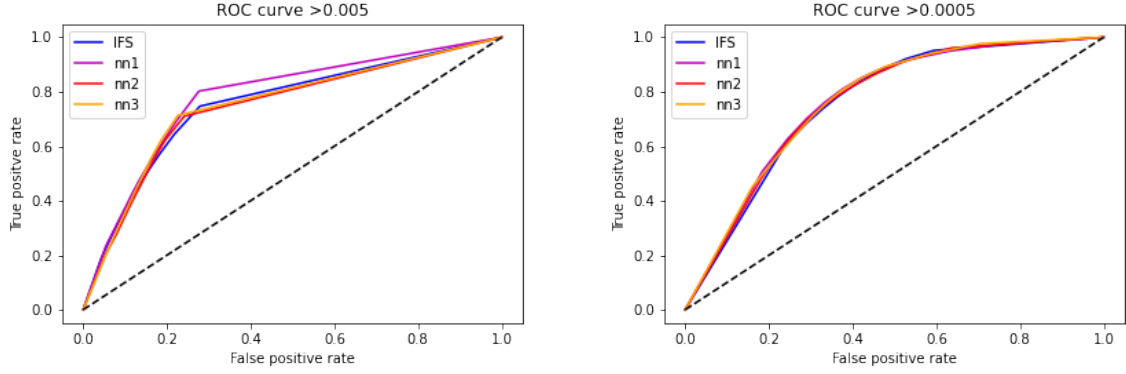


Figure 4. The ROC curves for the IFS and generated ensembles. The curves are computed after 2 days. Left for 0.005 and right for 0.0005 precipitation thresholds. The dashed diagonal line represents random chance, such that only points above it are a meaningful result.

4.2 CRPS

In Fig. 3 we show the CRPS value after 0, 1, 2 and 3 days averaged over the testing dataset. Here, we take the first step of the control run started after 0, 1, 2 and 3 days as a reference and compute the precipitation at those times based on the IFS ensemble and generated ensemble. We see in Fig. 3 that the CRPS values are almost the same, also for the different resolutions.

4.3 ROC curve

We use the data from the control run started after 2 days as the reference and compute the precipitation of the IFS and generated ensembles. We subtract the total precipitation at step 8 (day 2) of the step 4 (day 1). Then, we set all values of the reference below the thresholds (0.005 and 0.0005) to 0 and above to 1 to create a binary classification. We do the same for each of the 10 ensemble member to generate a probability for the specific rain rate. Then, we compute the mean of the ROC curves over the year 2022. In Fig. 4 we see that the ROC curves of the generated ensembles agree with the one produced by the IFS ensemble.

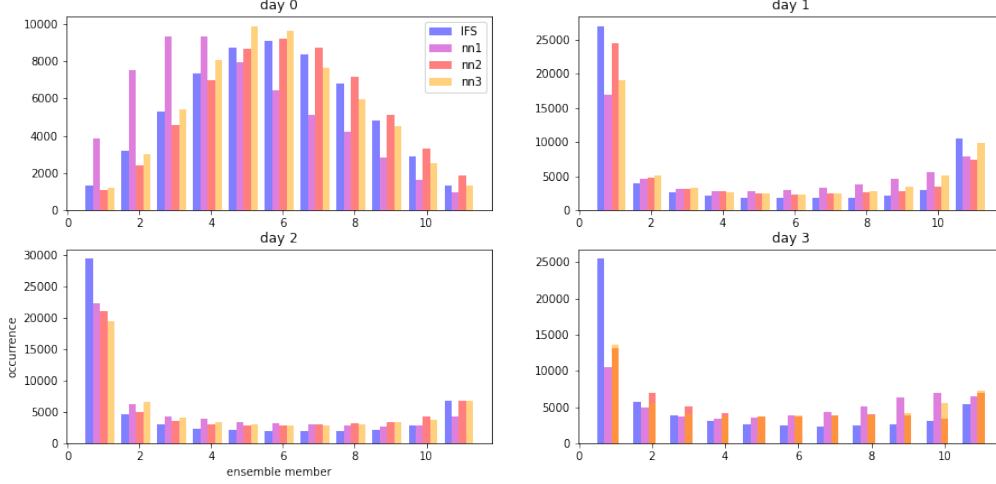


Figure 5. Comparison of the rank histograms for the generated and IFS ensemble after 0, 1, 2 and 3 days with the control run step 0 at those days as a reference.

4.4 Rank histogram

With the rank histogram we want to evaluate if the generated ensemble has a similar calibrated ensemble spread compared to the IFS ensemble. We use the data from the first control run step which is started after 0, 1, 2 and 3 days as the reference. Then we compute the difference with the previous day for the IFS and generated ensembles to get the same precipitation forecast. In Fig. 5 we observe that the ensemble spread of the generated and IFS ensemble are very similar. The non uniform shape of the histogram could be related to using the control run started after day 0, 1, 2 and 3 as a reference. However, here we are only interested in the similarity of the ensemble.

4.5 Case study

To present a case study, we evaluate the precipitation ensemble members at different cities and compare them with the control run on three different randomly selected days, January 15th, April 25th and August 3rd. In Figs. 6 and 7 we show the ensemble generated from the neural network *nn2*, which was trained on the degraded data taking every second grid point. We note that the ensemble from the network *nn1* and *nn3* is qualitatively similar. In general the spread of the generated and IFS ensemble is very similar.

5 Conclusions

In this paper, we present a new method to generate ensemble members from one deterministic control run. We want to highlight that this method does not rely on any high resolution samples to generate ensemble members on higher resolution simulations. We trained a conditional generative adversarial network, an adaptation of the `pix2pix` model, for this purpose. Using different verification methods, we could demonstrate that the generated ensemble members have close statistical properties compared to the IFS ensemble.

A limitation of this work is that precipitation events are globally sparse, such that it is difficult to train one accurate global model with this setup. We also note that precipitation has

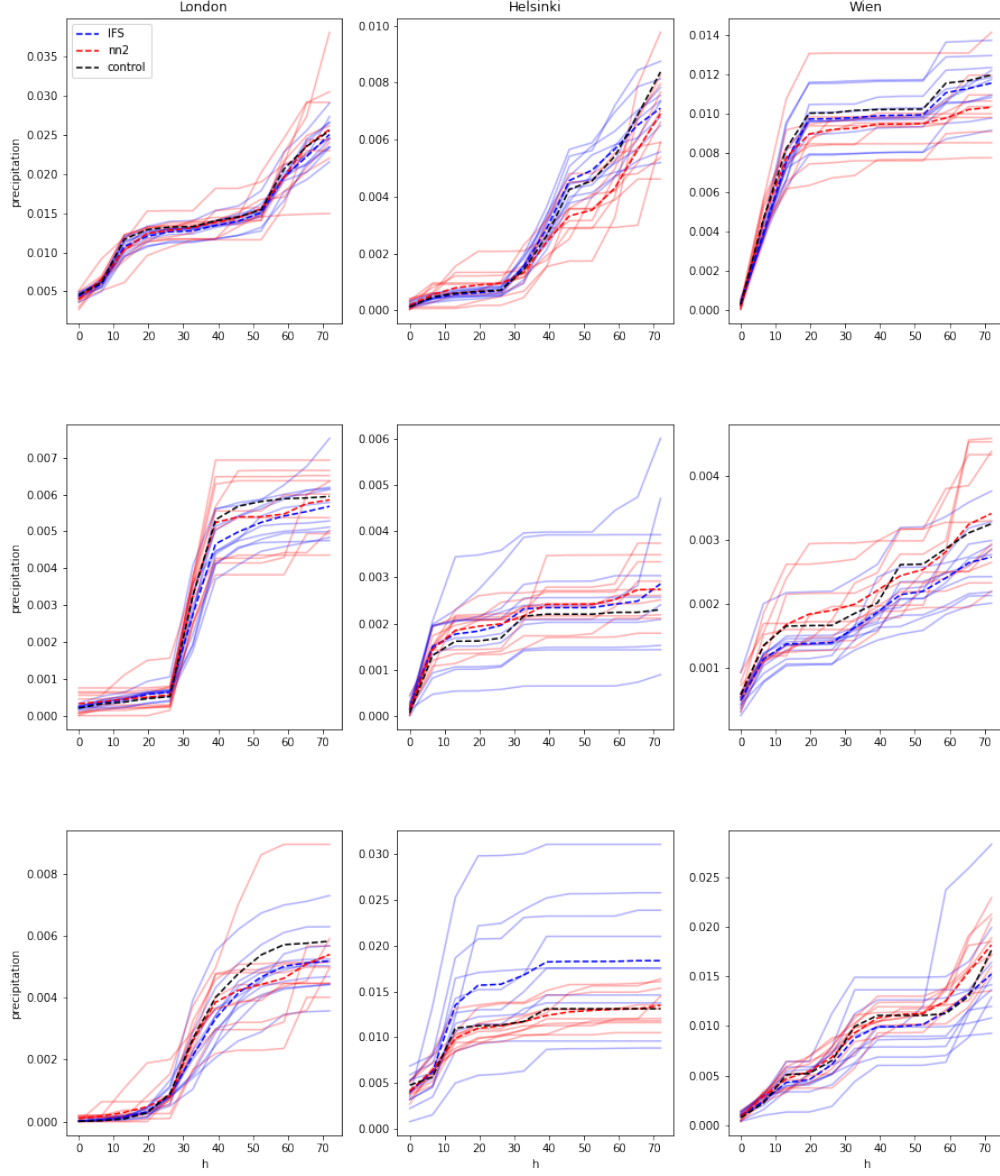
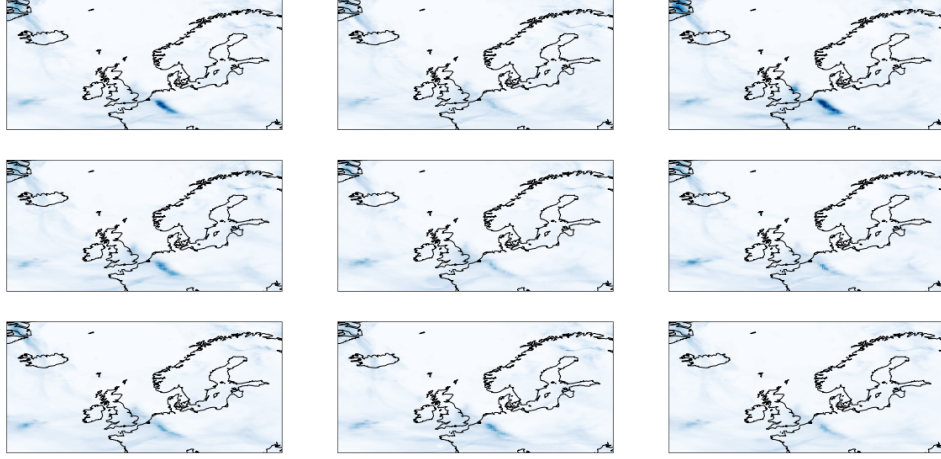
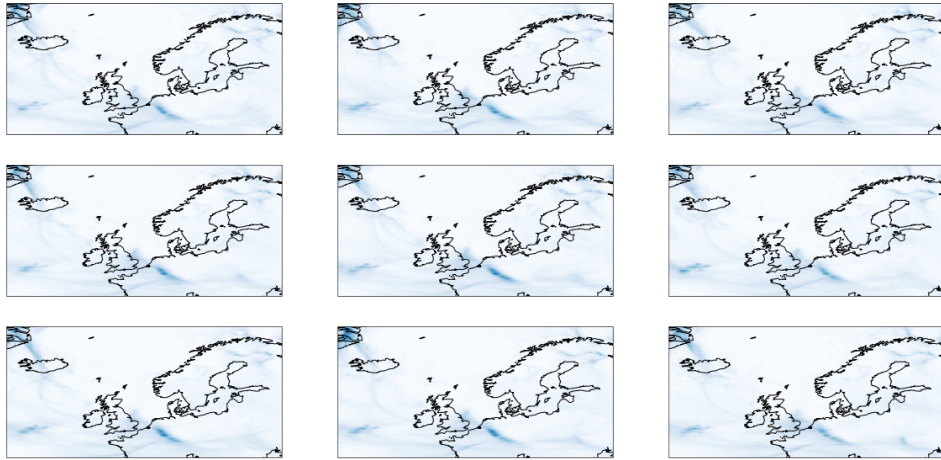


Figure 6. We show the IFS and generated ensemble members at different times of the year (January 15th: first row, April 25th: middle row, August 3rd: bottom row) in three different city. Here, we averaged the precipitation over a 10×10 pixel stencil around the city. The solid lines correspond to different ensemble members and the dashed line to their mean.



(a) Generated ensemble members using the neural network $nn2$.



(b) IFS ensemble members.

Figure 7. We show the IFS and generated ensemble members on 15th January using $nn2$. The results for April 25th and August 3rd, and for $nn3$, are both qualitatively and quantitatively similar and thus not shown here.

different characteristics in the tropics and extra-tropics, and data quality varies based on spatial locations [24]. As a consequence, a new set of neural networks needs to be trained for each region of interest. Given the relatively short training times of the proposed neural networks and the continuous increase in dedicated artificial intelligence hardware, we do not believe though that this is a serious issue. Additionally, many practical applications only require accurate precipitations forecasts over limited spatio-temporal domains.

There are various possible extensions of this work. To improve the accuracy of predicting extreme precipitation events, utilizing more training data, and potentially more powerful neural network architectures, could be beneficial. In this paper we have used generative adversarial networks as the main architecture. It is well-known that these can suffer from training difficulties such as mode collapse leading to less diverse generated samples compared to architectures based on, e.g., diffusion models [19]. While we have not observed mode collapse in practice, and the generated ensembles have excellent statistical properties, for the further improvement of results, it may be beneficial to look into more advanced image-to-image translation architectures.

We also anticipate that our model has the potential to be extended for broader application in different geographical regions, and for different meteorological parameters. It will also be interesting to see if the inclusion of additional input features, such as temperature, dew point, and wind speed, could further improve the obtainable results.

Acknowledgements

The authors thank Thomas Haiden and Mariana Clare for helpful discussion. This research was undertaken thanks to funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project-ID 274762653 - TRR 181, the Canada Research Chairs program and the NSERC Discovery Grant program. The authors also acknowledge support from the ECMWF special project *Mining 5th generation reanalysis data for changes in the global energy cycle and for estimation of forecast uncertainty growth with generative adversarial networks*.

References

- [1] Bihlo A., A generative adversarial network approach to (ensemble) weather prediction, *Neural Netw.* **139** (2021), 1–16.
- [2] Bihlo A. and Popovych R.O., Physics-informed neural networks for the shallow-water equations on the sphere, *J. Comput. Phys.* **456** (2022), 111024.
- [3] Brecht R. and Bihlo A., Computing the ensemble spread from deterministic weather predictions using conditional generative adversarial networks, *Geophys. Res. Lett.* **50** (2023), e2022GL101452.
- [4] Clare M., Jamil O. and Morcrette C.J., Combining distribution-based neural networks to predict weather forecast probabilities, *Q. J. R. Meteorol. Soc.* **147** (2021), 4337–4357.
- [5] Feser F., Rockel B., von Storch H., Winterfeldt J. and Zahn M., Regional climate models add value to global model data: a review and selected examples, *Bull. Am. Meteorol. Soc.* **92** (2011), 1181–1192.
- [6] Geer A.J., Learning earth system models from observations: machine learning or data assimilation?, *Phil. Trans. R. Soc. A* **379** (2021), 20200089.

- [7] Gentine P., Pritchard M., Rasp S., Reinaudi G. and Yacalis G., Could machine learning break the convection parameterization deadlock?, *Geophys. Res. Lett.* **45** (2018), 5742–5751.
- [8] Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A. and Bengio Y., Generative adversarial nets, in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.
- [9] Grönquist P., Ben-Nun T., Dryden N., Dueben P., Lavarini L., Li S. and Hoefler T., Predicting weather uncertainty with deep convnets, *arXiv preprint arXiv:1911.00630* (2019).
- [10] Grönquist P., Yao C., Ben-Nun T., Dryden N., Dueben P., Li S. and Hoefler T., Deep learning for post-processing ensemble weather forecasts, *Philos. Trans. R. Soc. A* **379** (2021), 20200092.
- [11] Hamill T.M., Interpretation of rank histograms for verifying ensemble forecasts, *Mon. Weather Rev.* **129** (2001), 550–560.
- [12] Harris L., McRae A.T., Chantry M., Dueben P.D. and Palmer T.N., A generative deep learning approach to stochastic downscaling of precipitation forecasts, *J. Adv. Model. Earth Syst.* **14** (2022), e2022MS003120.
- [13] Hersbach H., Decomposition of the continuous ranked probability score for ensemble prediction systems, *Weather Forecast.* **15** (2000), 559–570.
- [14] Hess P., Drüke M., Petri S., Strnad F.M. and Boers N., Physically constrained generative adversarial networks for improving precipitation fields from Earth system models, *Nat. Mach. Intell* **4** (2022), 828–839.
- [15] Holden Z.A., Abatzoglou J.T., Luce C.H. and Baggett L.S., Empirical downscaling of daily minimum air temperature at very fine resolutions in complex terrain, *Agric. For. Meteorol.* **151** (2011), 1066–1073.
- [16] Isola P., Zhu J.Y., Zhou T. and Efros A.A., Image-to-image translation with conditional adversarial networks, in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1125–1134.
- [17] Krizhevsky A., Sutskever I. and Hinton G.E., Imagenet classification with deep convolutional neural networks, in *Advances in Neural Information Processing Systems*, vol. 25, Curran Associates, 2012, pp. 1097–1105.
- [18] Leutbecher M. and Palmer T.N., Ensemble forecasting, *J. Comput. Phys.* **227** (2008), 3515–3539.
- [19] Li H., Yang Y., Chang M., Chen S., Feng H., Xu Z., Li Q. and Chen Y., Srdiff: Single image super-resolution with diffusion probabilistic models, *Neurocomputing* **479** (2022), 47–59.
- [20] Pathak J., Subramanian S., Harrington P., Raja S., Chattopadhyay A., Mardani M. and Kurth T., Hall D., Li Z., Azizzadenesheli K. *et al.*, Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators, *arXiv preprint arXiv:2202.11214* (2022).
- [21] Roebber P.J. and Bosart L.F., The complex relationship between forecast skill and forecast value: A real-world analysis, *Weather Forecast.* **11** (1996), 544–559.

- [22] Scher S. and Messori G., Ensemble methods for neural network-based weather forecasts, *J. Adv. Model. Earth Syst.* **13** (2021).
- [23] Shi X., Chen Z., Wang H., Yeung D.Y., Wong W.k. and Woo W.c., Convolutional LSTM network: a machine learning approach for precipitation nowcasting, in *Advances in Neural Information Processing Systems*, vol. 28, Curran Associates, 2015, pp. 802–810, arXiv:1506.04214.
- [24] Villarini G., Mandapaka P.V., Krajewski W.F. and Moore R.J., Rainfall and sampling uncertainties: A rain gauge perspective, *J. Geophys. Res. Atmos.* **113** (2008).
- [25] Weyn J.A., Durran D.R. and Caruana R., Can machines learn to predict weather? Using deep learning to predict gridded 500-hPa geopotential height from historical weather data, *J. Adv. Model. Earth Syst.* **11** (2019), 2680–2693.
- [26] Weyn J.A., Durran D.R., Caruana R. and Cresswell-Clay N., Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models, *J. Adv. Model. Earth Syst.* **13** (2021), e2021MS002502.
- [27] Wilks D.S., *Statistical methods in the atmospheric sciences*, vol. 100, Academic press, 2011.
- [28] Yu W., Nakakita E., Kim S. and Yamaguchi K., Impact assessment of uncertainty propagation of ensemble nwp rainfall to flood forecasting with catchment scale, *Adv. Meteorol.* **2016** (2016).