

Stay Report

To: GeoSphere, Vienna, Austria

Period: October 6th – October 31st, 2025

Topic: Work on machine learning post-processing methods

Supervisor: Irene Schicker, PhD

Chance of rain NWP postprocessing using neural networks

1. Introduction

During my four-week stay at GeoSphere, I worked on machine learning (ML) post-processing methods for generating point-based probabilistic precipitation forecasts from numerical weather prediction (NWP) model output. Machine learning methods, particularly neural networks, have gained considerable popularity due to their high-performing capabilities, a wide variety of situations where they are applicable, and relatively simple implementation. The main objective of this research is to evaluate and compare probabilistic precipitation forecasts generated by neural network-based post-processing of deterministic NWP models (ALADIN and IFS) with IFS ensemble forecasts. Furthermore, the transferability of the ML models between different NWP systems and geographical domains will be investigated.

Post-processing of NWP outputs is a way to further improve forecast performance. A more challenging approach to forecast improvement is to post-process the entire gridded forecast fields. For example, Xiang et al. (2022) used the Spatio-Temporal Transformer U-Net to perform bias correction and temporal downscaling for different meteorological variables, while Liu et al. (2024) used a Transformer based model to improve precipitation forecasts. The input NWP models used in such studies are often global (e.g. Price and Rasp 2022), and many studies focus on ensemble forecasts (e.g. Hess and Boers 2022). A more traditional approach is station-based post-processing, where forecasts are corrected only at locations with available observations. For example, the analog method is one such method in which using similarity search and the observation from training improves the forecast by exploiting historical forecasts with similar atmospheric conditions. Improvements in forecast performance (e.g. Vujec and Odak Plenković 2023; Odak Plenković et al. 2020) have been demonstrated for various variables (e.g. 2-m temperature, 10-m wind speed). Such methods also exist for the precipitation post-processing (Rojas-Campos et al. 2023). Although transfer learning and model generalization have attracted increasing attention within the context of weather forecast post-processing, relatively few studies have investigated the transferability of neural-network-based post-processing models across geographical domains or between different NWP systems. In particular, this project will investigate two aspects of transferability: (i) whether a neural network trained for one geographical domain can be applied to another domain, and (ii) whether a model trained using one deterministic NWP model can be applied to forecasts produced by another NWP system. In addition, blending deterministic ALADIN and IFS forecasts within a neural-network framework will also be investigated. The target variable in this study is the probability of precipitation occurrence.

2. Data, algorithms and training principles

Dataset

A one-year dataset of NWP forecasts and corresponding observations was used in this study. The NWP models used were the operational 2-km Croatian ALADIN model and the operational ECMWF Integrated Forecasting System (IFS). Operational IFS ensemble forecasts were used as a benchmark for comparison. The testing period was September 2024, while the remaining 2024 data were split into training and validation sets in a 3:1 ratio.

The forecast range of the ALADIN model is 72 hours, whereas the blending configuration extends the forecast range to 240 hours. Hourly forecasts are available up to forecast hour 90, 3-hourly forecasts up to forecast hour 144, and 6-hourly forecasts thereafter. Only the 00 UTC model runs were considered in this research.

This ALADIN model variation is based on the non-hydrostatic dynamical core with 87 vertical levels, where the lowest level is 10m. For the analysis, only the closest grid point to the measurement site is used.

Observational data consisted of hourly precipitation measurements from Croatian and Austrian stations, which were used for both training and testing. The Croatian dataset includes 58 stations distributed across the country, while the Austrian dataset comprises 25 stations.

Neural networks

Deep learning is a subset of machine learning where artificial neural networks, algorithms inspired by the human brain, learn from large amounts of data. It is the part of supervised learning, used for both regression and classification. An example of simple feed-forward neural network is shown in Figure 1.

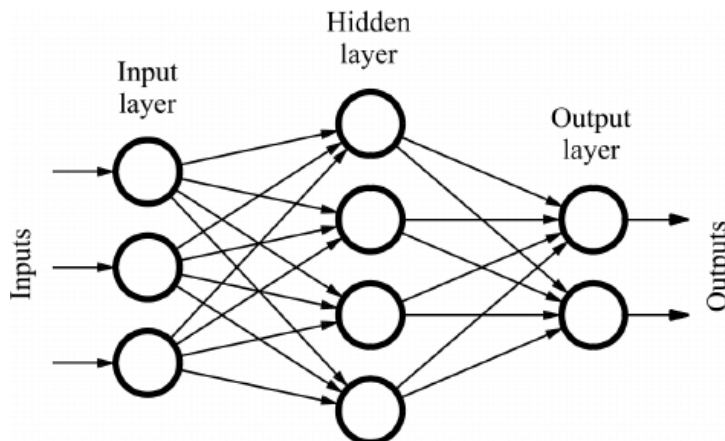


Figure 1. The example of simple feed-forward neural network

Neural networks are sets of algorithms intended to recognize patterns and interpret data. A neural network consists of connected units or nodes called neurons, which loosely model the neurons in a brain. These are connected by edges, which model the synapses in a brain. Each neuron receives signals from connected neurons, then processes them and sends a signal to other connected neurons. The "signal" is a real number, and the output of each neuron is computed by some non-linear function of the sum of its inputs, called the activation function. The strength of the signal at each connection is determined by a weight, which adjusts during the learning process.

Neurons are aggregated into layers. The signal travels from the first layer (the input layer) to the last layer (the output layer), possibly passing through multiple intermediate layers (hidden layers). A network is typically called a deep neural network if it has at least 1 hidden layer.

Training principles and tools

Neural networks are algorithms that can be highly sensitive to their architecture. In this study, no advanced hyperparameter optimization techniques were employed. Instead, conclusions regarding the optimal hyperparameter configuration (e.g., learning rate, L2 regularization strength, and network size) were drawn through an iterative trial-and-error process, supported by experiment tracking and analysis using the Weights & Biases platform. An example of the training process monitored through Weights & Biases is shown in Figure 2.



Figure 2. The example of the Weights and Biases platform

Since the networks ultimately exhibited relatively similar performance across different choices of activation function, learning rate, L2 regularization strength, and dropout rate, most hyperparameter tuning efforts focused on the network architecture, specifically its depth and width. The final configuration employed the ReLU activation function, a learning rate of 0.001, an L2 regularization coefficient of 0.001, and no dropout. The network architectures evaluated were: ($d = 2$, $2 = 5$), ($d = 4$, $w = 20$), ($d = 6$, $w = 40$), ($d = 8$, $w = 60$), and ($d = 10$, $w = 80$). Although several verification metrics were monitored during training, the validation loss, combined with early stopping, was used as the primary optimization objective.

3. Results

As previously mentioned, a one-year dataset was used for model development and evaluation. September 2024 was reserved as the test dataset for all experiments, while the remaining data were divided into training and validation sets using a 3:1 ratio.

The event was defined as precipitation exceeding 0.1 mm during the last hour, and the predictand was the probability of precipitation occurrence.

The analysis presented in this study consists of two main components: transferability and blending using one or more NWP model inputs. The transferability analysis is further divided into two parts: transferability across different geographical domains and transferability across different NWP models.

Transferability

The set of NWP predictors used in these experiments was relatively limited and consisted of 2-m temperature, 10-m wind speed and direction, and accumulated precipitation. Additional predictors included the time of day, the day of the year, and the forecast day.

The objective of this part of the study was to investigate whether a neural network model trained using output from the same NWP model (IFS in this case) could be successfully transferred to a different geographical domain. Specifically, a model trained using IFS forecasts and observations from the Austrian station network was evaluated using IFS forecasts and observations from the Croatian station network.

The first step was to identify the optimal neural network architecture using IFS forecasts and observations from the Austrian station network. Figure 3 presents the performance of different neural network architectures on the Austrian test dataset. For each architecture, one representative forecast was selected, and the corresponding verification results are shown.

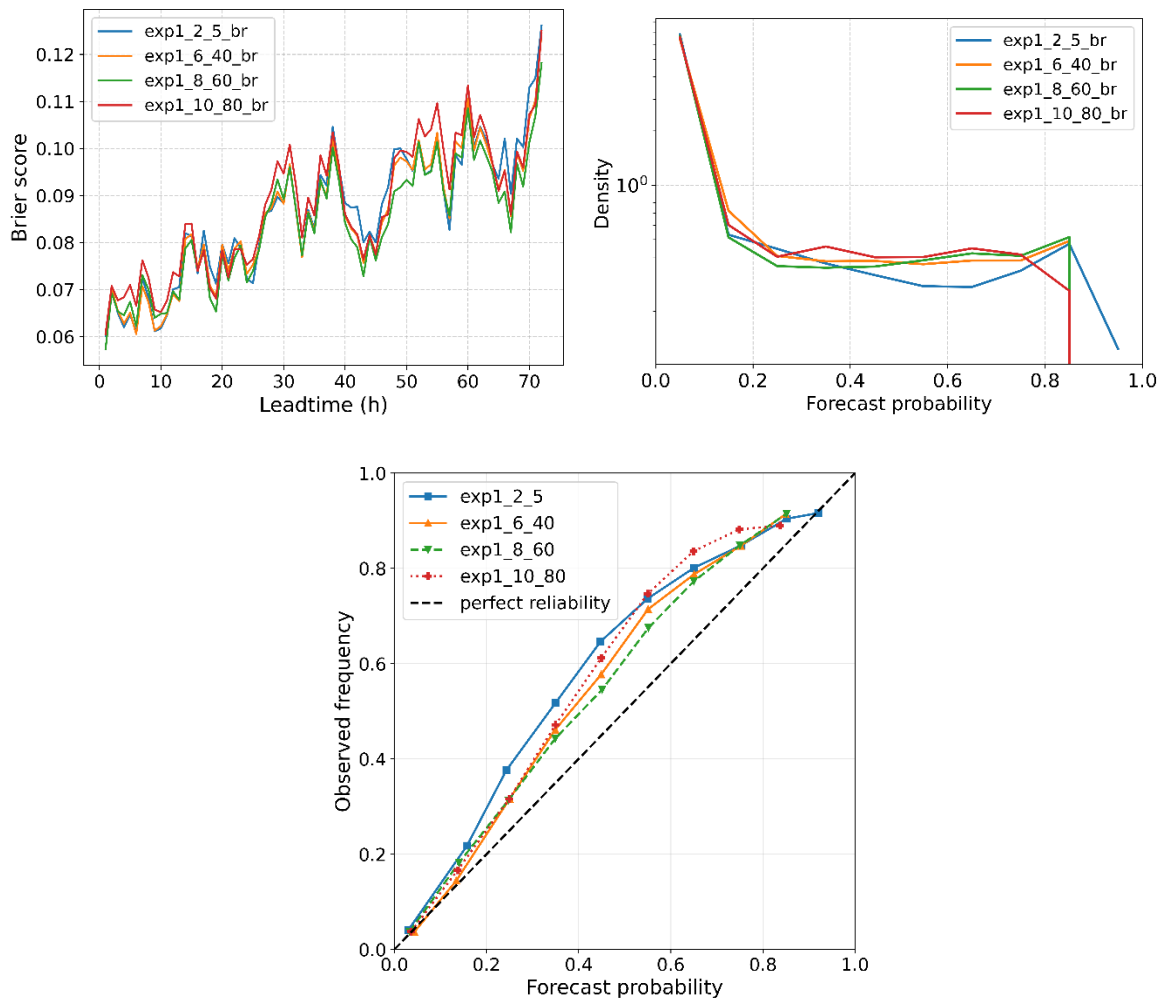


Figure 3. Brier score (upper left), frequency diagram (upper right), and reliability diagram (bottom) for probabilistic precipitation forecasts generated by different neural network architectures during the September 2024 test period using observations from the Austrian station network and IFS as NN input. The models were trained using IFS forecasts and observations from the Austrian station network.

Although most of the forecasts perform similarly, some differences can be observed. Overall, the largest network ($d = 10$, $w = 80$) generally exhibits the highest Brier scores, together with relatively poor reliability for forecast probabilities above 50% and comparatively low sharpness. On the other

hand, it shows better reliability than most of the other architectures for lower forecast probabilities (below 50%).

In contrast, the smallest network ($d = 2$, $w = 5$) shows mixed performance in terms of the Brier score. For some lead times, it achieves the lowest Brier scores, whereas for others it ranks among the poorest-performing models. Although it is less reliable for moderate forecast probabilities (20–60%), it performs well for high forecast probabilities (above 80%) and produces relatively sharp forecasts compared with the other architectures.

Based on all results, the architectures with dimensions ($d = 2$, $w = 5$) and ($d = 6$, $w = 40$) were selected for the subsequent analyses, as they represent a reasonable compromise between model complexity and predictive performance.

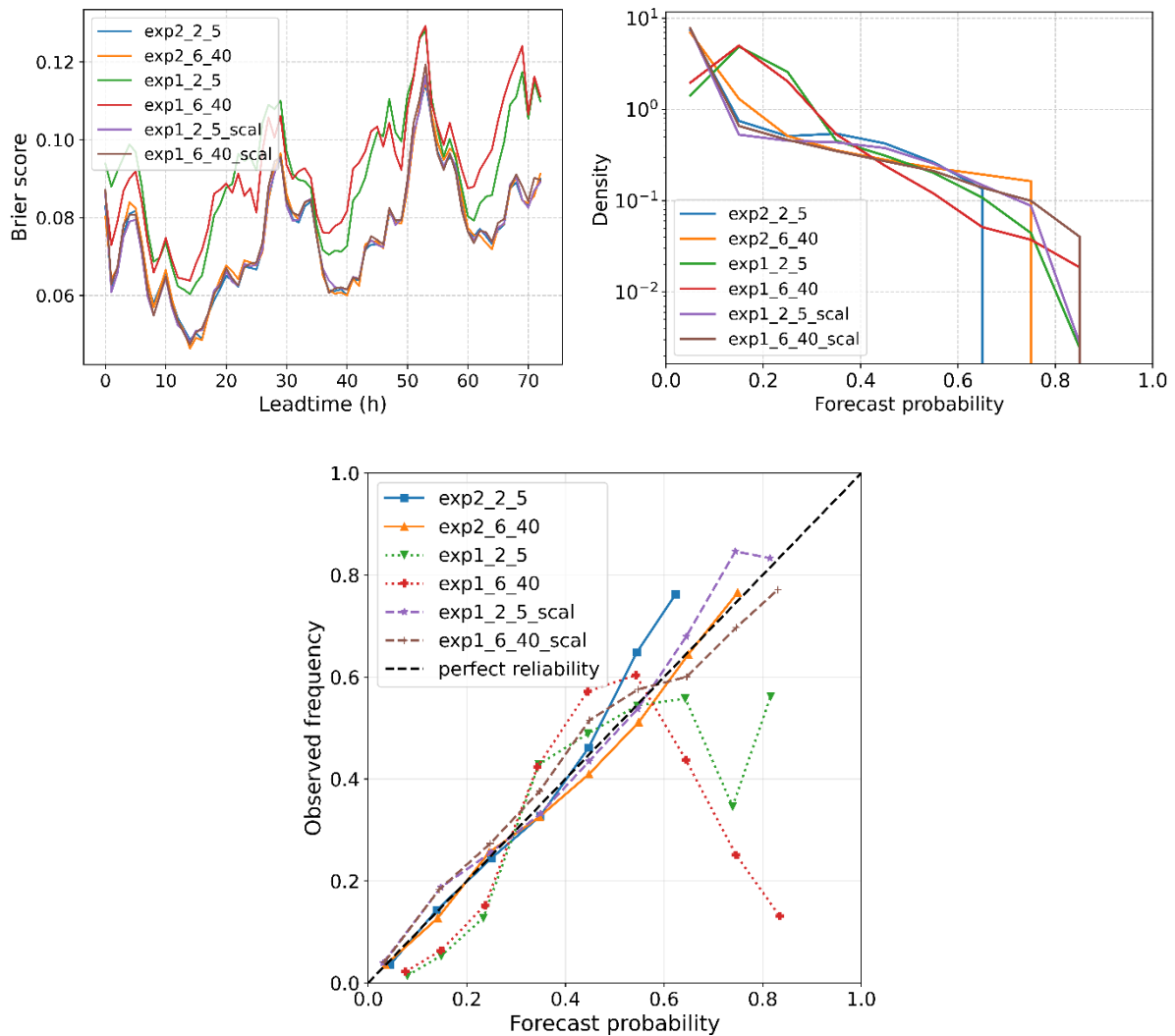


Figure 4. Brier score (upper left), frequency diagram (upper right), and reliability diagram (bottom) for probabilistic precipitation forecasts generated by different neural network architectures during the September 2024 test period using observations from the Croatian station network and IFS model as NN input. The models labelled „exp2“ were trained using IFS forecasts and observations from the Croatian station network, whereas all other models were trained using IFS forecasts and observations from the Austrian station network. The forecasts labelled scal use the scaler derived from the original training dataset (IFS forecasts and Austrian observations), while the remaining forecasts use scalers fitted to training datasets based on IFS forecasts and Croatian observations.

After selecting the best-performing architectures on the Austrian dataset, the first transferability experiment was conducted. The results of the transferability analysis across different geographical domains are presented in Figure 4.

The forecasts *exp1_2_5* and *exp1_6_40* exhibit noticeably poorer performance than the remaining models across all verification metrics. These forecasts correspond to neural networks trained using IFS forecasts and observations from the Austrian station network and evaluated using IFS forecasts together with observations from the Croatian station network. In these experiments, the models were applied using the scaler fitted to the Croatian dataset rather than the scaler used during the original model training. This mismatch in feature scaling considerably degraded forecast performance.

In contrast, the models trained on the Croatian dataset, as well as the transferred models using the original scaler from the Austrian training dataset, show comparable performance. Although some differences remain, the overall results suggest that transferability across geographical domains is feasible. These results suggest that a neural network model trained by one meteorological service could potentially be applied to another geographical domain, provided that forecasts from the same NWP model are available and the original preprocessing (including feature scaling) is preserved.

The objective of the next part of the study was to investigate the transferability of neural network models across both geographical domains and NWP models. Three models were considered: one trained using IFS forecasts and observations from the Austrian station network, one trained using IFS forecasts and observations from the Croatian station network, and one trained using ALADIN forecasts and Croatian observations. All three models were evaluated using ALADIN forecasts and observations from the Croatian station network. This experimental setup made it possible to assess the combined effects of domain and NWP-model transferability.

Figure 5. illustrates the importance of using the appropriate feature scaler. It is evident that both models trained using IFS forecasts and observations from the Austrian station network and those trained using IFS forecasts and observations from the Croatian station network (labelled *_scal*) achieve substantially lower Brier scores than the corresponding forecasts using the scaler fitted during ALADIN model training. Therefore, only the forecasts using the original training scaler were considered in the subsequent analysis.

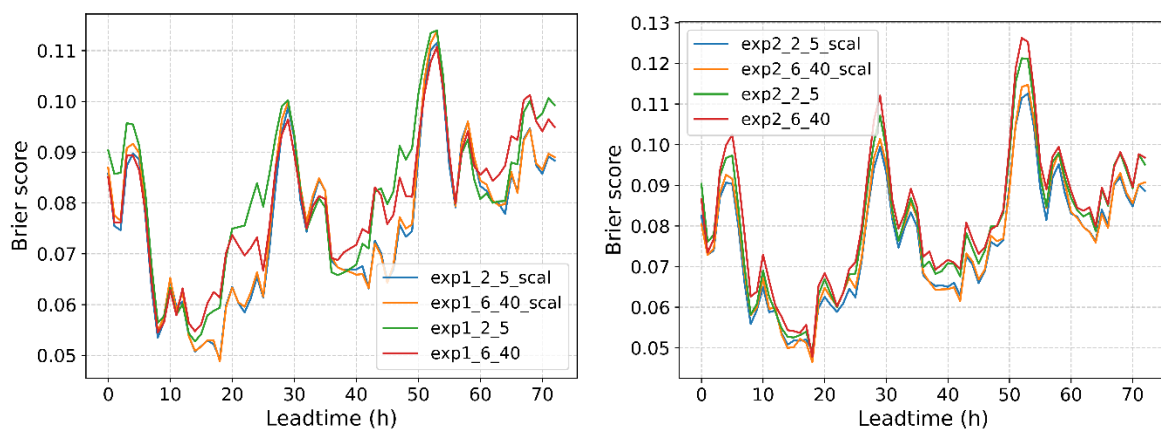


Figure 5. Brier score of different forecasts generated by the models that were trained using the IFS forecasts and observations from the Austrian station network (left), and using the IFS forecasts and observations from the Croatian station network (right) during the September 2024 test period using observations from the Croatian station network and ALADIN as NN input. The forecasts labelled *scal* use the scaler derived from corresponding training dataset, while the remaining forecasts use scalers fitted to training datasets based on ALADIN forecasts and Croatian observations.

Finally, the transferability results are presented in Figure 6. All three models exhibit very similar performance in terms of the Brier score and reliability. The model trained using ALADIN forecasts and Croatian observations achieves slightly better performance in both the Brier score and the reliability diagram than the other models; however, the differences are small and the overall results remain comparable. On the other hand, the model trained using IFS forecasts and observations from the Austrian station network produces the sharpest forecasts.

The results suggest that both domain and NWP-model transferability are feasible, with all investigated approaches achieving comparable overall performance. At a time when GPU resources may be limited, model transferability offers the possibility of applying a neural network trained for one geographical domain to another, or even using forecasts from a different NWP model as input, while still maintaining comparable forecast skill.

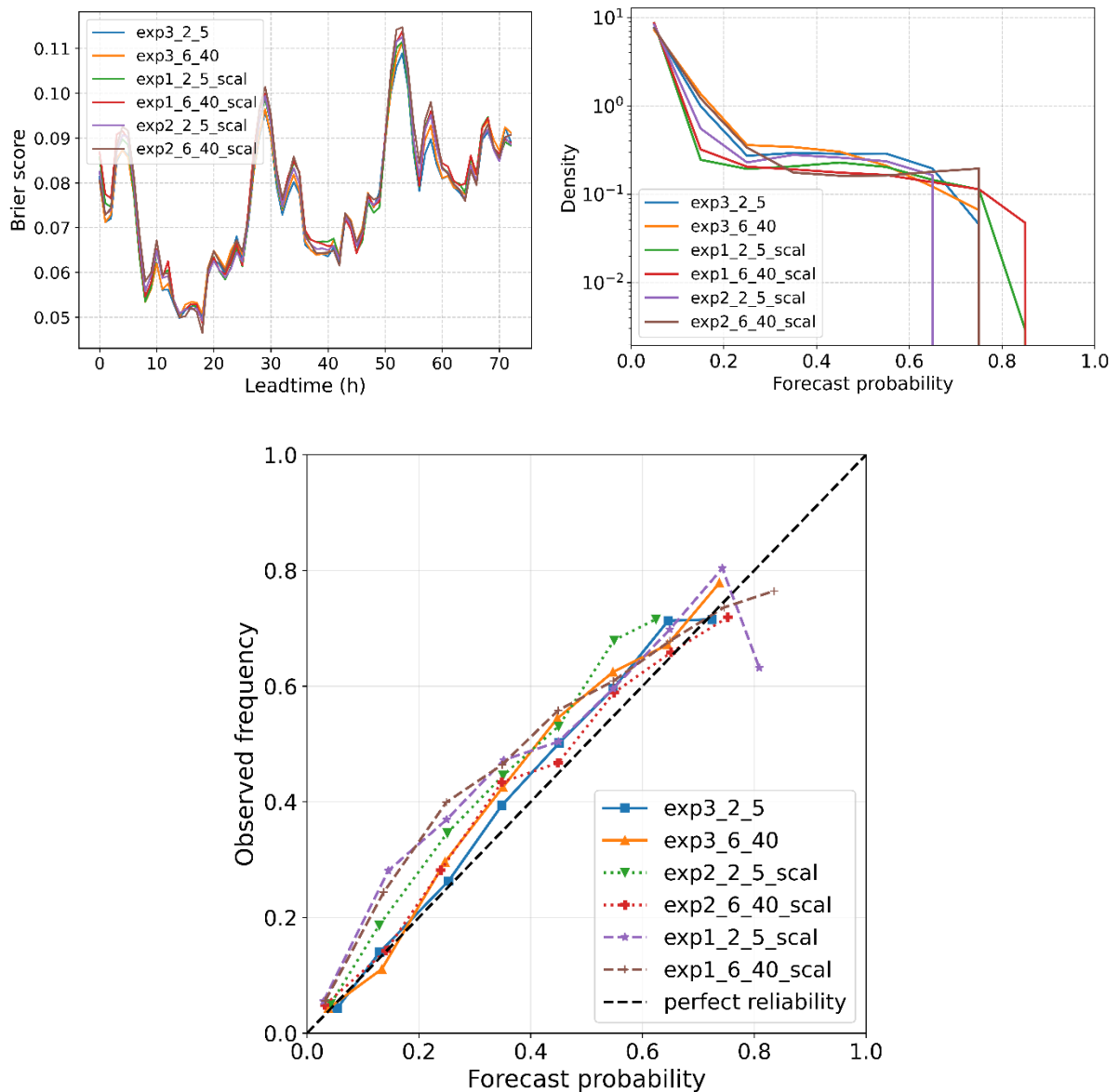


Figure 6. Brier score (upper left), frequency diagram (upper right), and reliability diagram (bottom) for probabilistic precipitation forecasts generated by different neural network architectures during the September 2024 test period using observations from the Croatian station network and ALADIN as NN input. The models labelled „exp1” were trained using IFS forecasts and observations from the Austrian station network, the models labelled „exp2” were trained using ALADIN forecasts and observations from the Croatian station network, whereas all other models were trained using ALADIN forecasts and observations from the Croatian station network. The forecasts labelled scal use the scaler derived from the corresponding

original training datasets, while the remaining forecasts use scalers fitted to training datasets based on ALADIN forecasts and Croatian observations.

Blending

The objective of this part of the study was to investigate the performance of a neural network model using information from both the deterministic ALADIN and IFS models. The performance of this approach was compared with that of the operational IFS ensemble prediction system. The Croatian dataset was used for all experiments in this part of the study.

The set of NWP predictors used in these experiments was expanded and consisted of the following ALADIN predictors: 2-m temperature, 10-m wind speed and direction, relative humidity, mean sea level pressure, 10-m wind gust, 80-m wind speed and direction, and accumulated precipitation. The IFS predictors included total precipitation, 10-m wind components, and 100-m wind components. Additional predictors included the accumulation interval, time of day, day of the year, and forecast day.

As described previously, the forecast range of the ALADIN model is 72 hours, whereas the blending configuration extends the forecast range to 240 hours. Hourly forecasts are available up to forecast hour 90, 3-hourly forecasts up to forecast hour 144, and 6-hourly forecasts thereafter. Only the 00 UTC model runs were considered in this study.

Since ALADIN forecasts are available only up to 72 hours, the last available value of each ALADIN predictor was propagated to the remaining forecast lead times. This persistence-based approach has been suggested to further improve forecast performance (Dabernig & Atencia, 2024).

The results suggest that all blending configurations outperform the operational IFS ensemble. All neural network forecasts exhibit lower Brier scores than the operational IFS ensemble for almost all forecast lead times and also demonstrate better reliability. On the other hand, the operational IFS ensemble produces sharper forecasts. It tends to assign high forecast probabilities more frequently than the neural network models; however, these forecasts are often overconfident, as they occur more frequently in the forecasts than in the corresponding observations.

Two neural network approaches were evaluated. Models labelled **34**, **36**, and **42** use predictor values only for the forecast hour of interest, whereas models labelled **56**, **59**, and **60** additionally include predictor values from the preceding and following forecast hours. Although all neural network models show comparable performance, the approach using only predictor values from the forecast hour of interest performs slightly better overall. These models achieve lower Brier scores and better reliability; however, they exhibit somewhat poorer sharpness, producing high forecast probabilities less frequently.

Overall, the results demonstrate the potential of neural networks to blend information from two deterministic NWP models for probabilistic precipitation forecasting. All neural network configurations outperform the operational IFS ensemble, while the differences between the individual neural network architectures remain relatively small.

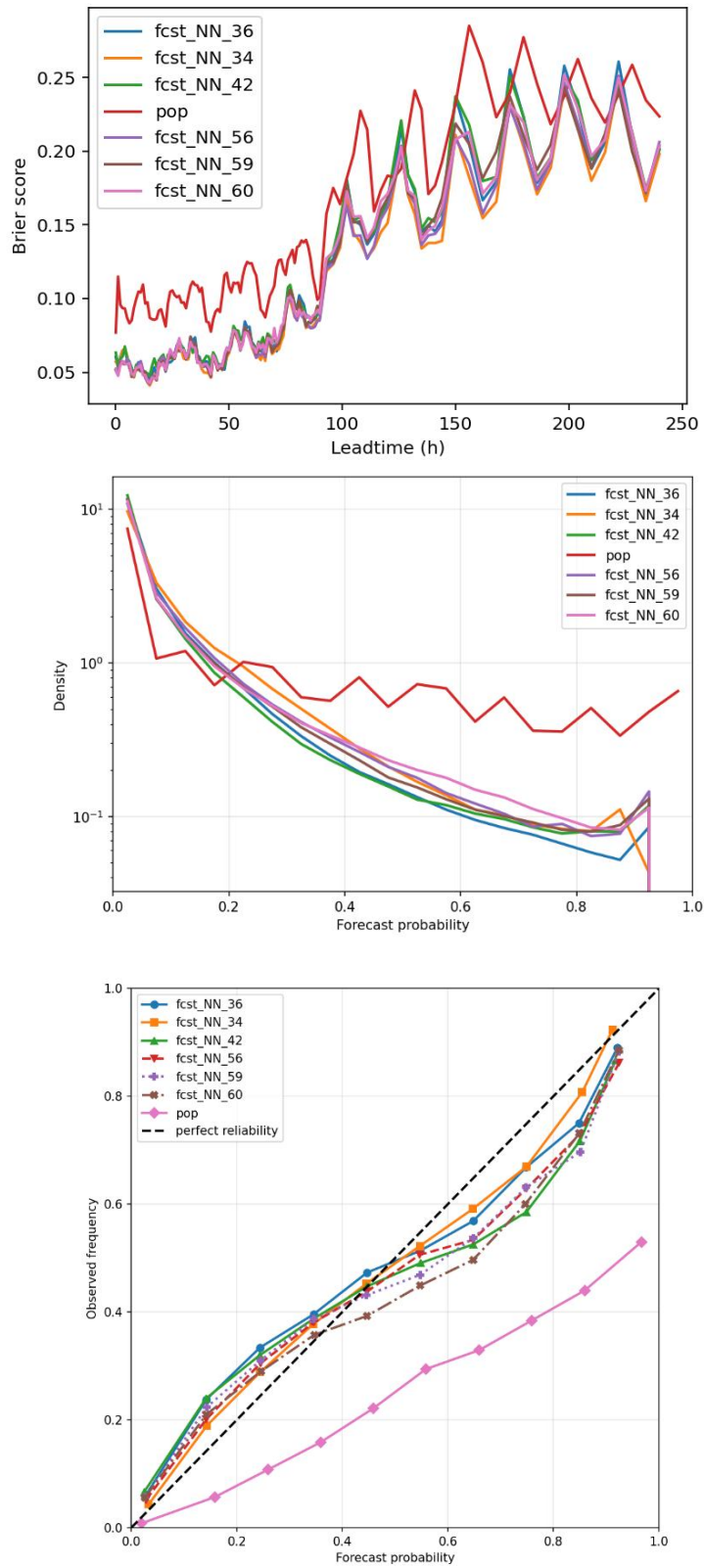


Figure 7. Brier score (top), frequency diagram (middle), and reliability diagram (bottom) for probabilistic precipitation forecasts generated by different neural network architectures during the September 2024 test period using observations from the Croatian station network and both ALADIN and IFS forecasts as neural network input. The models labelled **34**, **36**, and **42** use predictor values only for the forecast hour of interest, whereas models labelled **56**, **59**, and **60** additionally include predictor values from the preceding and following forecast hours. The **pop** refers to the operational IFS ensemble forecast.

4. Conclusion

The results of the transferability analysis suggest that both domain and NWP-model transferability are feasible, with all investigated approaches achieving comparable overall performance. These findings indicate that a neural network trained for one geographical domain or one NWP model can be successfully applied to another while maintaining comparable forecast skill. Such transferability could reduce the need for retraining machine learning models for every new domain or forecasting system. Consequently, meteorological services could share and reuse trained models, allowing valuable computational and GPU resources to be redirected towards other research and operational tasks.

Furthermore, the results demonstrate the potential of blending information from the deterministic ALADIN and IFS NWP models within a neural network framework for probabilistic precipitation forecasting. All investigated neural network configurations outperform the operational IFS ensemble, while the differences among the individual network architectures remain relatively small. These findings suggest that the proposed blending approach can provide more accurate and reliable probabilistic precipitation forecasts than the current operational IFS ensemble.

5. References

- Dabernig, M., & Atencia, A. (2024). "The Simplest Idea One Can Have" for Seamless Forecasts with Postprocessing. *arXiv preprint arXiv:2410.11916*. <https://doi.org/10.48550/arXiv.2410.11916>
- Hess, P., & Boers, N. (2022). Deep learning for Improving Numerical Weather Prediction of Heavy Rainfall. *Journal of Advances in Modeling Earth Systems*, 14, e2021MS002765. <https://doi.org/10.1029/2021MS002765>
- Liu, H., Fung, J. C. H., Lau, A. K. H., & Li, Z. (2024). Enhancing Quantitative Precipitation Estimation of NWP Model With Fundamental Meteorological Variables and Transformer Based Deep Learning Model. *Earth and Space Science*, 11, e2023EA003234. <https://doi.org/10.1029/2023EA003234>
- Odak Plenković, I., Schicker, I., Dabernig, M., Horvath, K., & Keresturi, E. (2020). Analog-based post-processing of the ALADIN-LAEF ensemble predictions in complex terrain. *Q J R Meteorol Soc* 146:1842–1860. <https://doi.org/10.1002/qj.3769>
- Price, I., & Rasp, S. (2022). Increasing the accuracy and resolution of precipitation forecasts using deep generative models. In *International conference on artificial intelligence and statistics* (pp. 10555-10571). PMLR. <https://doi.org/10.48550/arXiv.2203.12297>
- Rojas-Campos, A., Wittenbrink, M., Nieters, P., Schaffernicht, E. J., Keller, J. D., & Pipa, G. (2023). Postprocessing of NWP Precipitation Forecasts Using Deep Learning. *Weather and Forecasting*, 38(3), 487-497. <https://doi.org/10.1175/WAF-D-21-0207.1>
- Vujec, I., & Odak Plenković, I. (2023). Kalman filter sensitivity tests for the NWP and analog-based forecasts post-processing. *Meteorol Atmos Phys* 135, 1. <https://doi.org/10.1007/s00703-022-00939-w>
- Xiang, L., Guan, J., Xiang, J., Zhang, L., & Zhang, F. (2022). Spatiotemporal model based on transformer for bias correction and temporal downscaling of forecasts. *Front. Environ. Sci.* 10:1039764. <https://doi.org/10.3389/fenvs.2022.1039764>