A calibrated and consistent combination of probabilistic forecasts for the exceedance of several precipitation thresholds using neural networks.

Peter Schaumann¹, Reinhold Hess², Martin Rempel², Ulrich Blahak², and Volker Schmidt¹

¹Institute of Stochastics, Ulm University, Ulm, Germany (peter.schaumann@uni-ulm.de)
²Deutscher Wetterdienst, Offenbach, Germany

In this talk we present a new statistical method for the seamless combination of two different ensemble precipitation forecasts (Nowcasting and NWP) using neural networks (NNs), see [1]. The method generates probabilistic forecasts for the exceedance of a set of predetermined thresholds (from 0.1mm up to 5mm). The aim of the combination model is to produce seamless and calibrated forecasts which outperform both input forecasts for all lead times and which are consistent regarding the considered thresholds. First, the hyper-parameters of the NNs are chosen according to a certain hyper-parameter optimization algorithm (not to be confused with the training of the NNs itself) on a 3-month dataset (dataset A). Then, the resulting NNs are tested via a rolling origin validation scheme on two 3-month datasets (datasets B & C) with different input forecasts each. Datasets A & B contain forecasts of DWD’s RadVOR, a radar-based nowcasting system, and Ensemble-MOS, a post-processing system of NWP ensembles made by COSMO-DE-EPS, with a horizontal resolution of 20km, which is a predecessor of ICON-D2-EPS. Ensemble-MOS forecasts were provided for up to +6h, while RadVOR forecasts were available up to +2h. For dataset C, forecasts with a grid size of 2.2km are used from STEPS-DWD, a new implementation of the Short-term Ensemble Prediction System (STEPS) by DWD, and ICON-D2-EPS as a NWP ensemble system. Forecasts were made up to +6h. In both validation datasets (B & C), the forecasts show the well-known behavior that the nowcasting systems RadVOR & STEPS are superior for short lead times, while NWP forecasts (Ensemble-MOS & ICON-D2-EPS) outperform these systems for later lead times. Based on the comparison of several validation scores (bias, Brier skill score, reliability and reliability diagram) we can show that the combination is indeed calibrated, consistent and outperforms both input forecasts for all lead times. It should be noted that the combination works on dataset C, although the hyper-parameters were chosen based on dataset A, which contains different forecasts for a different grid size.