Adaptive Blending of Probabilistic Precipitation Forecasts with Emphasis on Calibration and Temporal Forecast Consistency

Martin Rempel[1,†], Peter Schaumann[2], Reinhold Hess[1], Volker Schmidt[2] and Ulrich Blahak[1]

[1]*Deutscher Wetterdienst, Offenbach, Germany*

[2]*Institute of Stochastics, Ulm University, Ulm, Germany*

Corresponding author: Martin Rempel, martin.rempel@dwd.de

MR and PS contributed equally to this work.

ABSTRACT: A wealth of forecasting models is available for operational weather forecasting. Their strengths often depend on the lead time considered, which generates the need for a seamless combination of different forecast methods. The combined and continuous products are made in order to retain or even enhance the forecast quality of the individual forecasts and to extend the lead time to potentially hazardous weather events. In this study, we further improve an artificial neural network based combination model that was recently proposed in a previous paper. This model combines two initial precipitation ensemble forecasts and produces exceedance probabilities for a set of thresholds for hourly precipitation amounts. Both initial forecasts perform differently well for different lead times, whereas the combined forecast is calibrated and outperforms both initial forecasts with respect to various validation scores and for all considered lead times (+1h to +6h). Moreover, the robustness of the combination model is tested by applying it to a new dataset and by evaluating the spatial and temporal consistency of its forecasts. The changes proposed further improve the forecast quality and make it more useful for practical applications. Temporal consistency of the combined product is evaluated using a flip-flop index. It is shown that the combination provides a higher persistence with decreasing lead times compared to both input systems.

1. Introduction

The term adaptive blending stands for the search of an optimal lead-time- and position-dependent weighting between two or more forecasts that cover different forecast ranges or are based on different forecast models, e.g., different configurations of numerical weather prediction (NWP). The term blending represents a part of the so-called seamless prediction, where this term was originally introduced by Palmer et al. (2008) to describe the combination of weather prediction and climate modeling as a unified topic. In the course of time, the definition of seamless prediction was extended to include also interactions with biogeophysical components (Hazeleger et al. 2012) and, more generally, to describe the interactions between weather, climate, and the Earth system (Brunet et al. 2010). A recent publication by Ruti et al. (2020) defines seamless prediction as a whole value cycle enclosing four parts: the generation of information, the dissemination to users, the perception and decision-making, and the outcomes and values.

The currently ongoing project SINFONY (Seamless INtegrated FOrecastiNg sYstem) of Deutscher Wetterdienst (DWD) can be assigned to the topic of seamless prediction, since it focuses on the seamless prediction of precipitation within the short-term range up to +12 h ahead. Here, the term seamless is referred to as the combination of forecasts of observation-based precipitation nowcasting techniques with those of the NWP. The goals of SINFONY to achieve this combination are twofold and address two parts of the definition of seamless prediction mentioned above. First, the generation of information is addressed by individually improving both forecasting methods — nowcasting and NWP — in such a way that in terms of a verification metric the gap between each of them is narrowed. Based on these improvements, the development and implementation of tailor-made combination methods will lead to a user-oriented unique forecast including condensed information of both individual forecasts. Moreover, the interaction with users is addressed by using the feedback of hydrological services and forecasters to further improve the products.

Nowcasting and NWP forecasts can provide valuable guidance for users on different lead time scales (Heizenreder et al. 2015; Hess 2020). Many common precipitation nowcasting methods rely on the Lagrangian persistence approach, whereby the latest field of observed reflectivities or estimated rain rates is extrapolated in space and time by a previously determined motion vector field (Germann and Zawadzki 2002). Due to this purely advective approach, the dynamic uncertainty induced by growth and decay of precipitation patterns is not considered. Thus, the quality of such forecasts is high as long as the Lagrangian persistence assumption is valid (Zawadzki et al. 1994). The prediction of specific weather events depends on their spatial extent (Venugopal et al. 1999) and can reach from minutes when considering small-scale phenomena (e.g. single thunderstorms) up to hours at length scales of several hundred kilometers (Foresti and Seed 2014).

The physical evolution of precipitation fields is, on the other hand, explicitly simulated by NWP models. However, one source of forecast errors of the latter can be found in initial and boundary conditions as well as in inexact solutions of approximated physical equations due to finite resolutions in time and space. Nicolis et al. (2009) showed that subgrid parameterizations of cloud microphysics are especially important for precipitation forecasting. However, shortcomings in such a parameterization lead to deficiencies in simulated rainfall intensities (Stephan et al. 2008).

Despite these error sources, NWP forecasts are able to outperform the forecast quality of precipitation nowcasting techniques 2-3 h after initialization, as it will be shown in Section 4. Therefore, the seamless combination aims to create a unique and consistent forecast in which the best skill is retained and the amount of information is condensed regardless of location and lead time (Brunet et al. 2015).

Vannitsem et al. (2021) present, among others, an overview of methods to combine forecasts of nowcasting and NWP and further point out that this combination may take place in physical or probability spaces. The weighting mentioned above can be based on a long-term comparative verification of both initial forecast systems. This is done in physical space by Golding (1998) in Nimrod, one of the first combination schemes, or in a probability space by Kober et al. (2012). Haiden et al. (2011) utilized in INCA (Integrated Nowcasting through Comprehensive Analysis) a simple linear weighting function, in which the weight for NWP forecasts increases from 0 at the beginning to 1 at a lead time of +4 h. The Short-Term Ensemble Prediction System (STEPS; Seed (2003), Seed et al. (2013)) in its implementation by Bowler et al. (2006) quantifies in real time not only tendencies in a sequence of the latest observations, but also the skill of a NWP forecast to adjust weights for combining the nowcast extrapolation and the NWP forecast, depending on lead time and spatial length scale. A forecast ensemble is then generated by replacing nonpredictable scales with spatial correlated random noise. Moreover, the emergence of nowcasting ensemble techniques allows to use the ensemble spread as an objective combination metric. Based on this, e.g. Nerini et al. (2019) implemented an ensemble Kalman filter for the iterative combination of

NWP forecasts and precipitation nowcasting extrapolations. Johnson and Wang (2012) as well as Bouttier and Marchal (2020) carried out combinations of multimodel ensembles.

With focus on nowcasting approaches based on machine-learning (ML) techniques, many studies use model information as an additional predictor. In Han et al. (2017), radar observations combined with data of the analysis system VDRAS (Variational Doppler Radar Analysis System) are utilized to train a support vector machine (SVM) for answering the question whether there will be reflectivity > 35 dBZ within a box in the next 30 min based on the information in the adjacent boxes. Ukkonen et al. (2017) utilize an artificial neural network (ANN) with lightning and reanalysis data as input to evaluate thunderstorm predictors for Finland. An overview about machine-learning approaches with focus on nowcasting is given by, e.g., Prudden et al. (2020) and Cuomo and Chandrasekar (2021).

Besides accuracy, calibration and spatial consistency, also temporal consistency is desired for operational forecasts. Here and in the following, the notion "temporal consistency" is not to be understood as the time-dependent correlation structure of a single forecast. Rather, it describes the variability between a number of model runs for a fixed valid time that is also often referred to as *jumpiness*. Ideally, there is a large uncertainty in early forecasts that decreases with time so that forecasts converge towards the observations and become more and more confident. However, in practice, it is often observed that updated forecasts for one specific time and location exhibit spurious jumps due to forecast errors. This is a problem for meteorologists, who want to rely on the most topical numerical forecast and may need to revise their opinion accordingly, especially in case of weather warnings. It seems to be very unreasonable if a warning is issued, canceled soon thereafter, and may be even re-issued again, see e. g. Griffiths et al. (2019).

In the present paper two forecast systems are combined (nowcasting and NWP); each one has its own characteristics in temporal consistency, which affect the consistency of the combined product. The transition from nowcasting to NWP with larger forecast lead time may likely result in additional inconsistencies, since the systematic errors of the two systems differ. Moreover, the method of combination itself may introduce additional inconsistencies, e. g. if individual architectures or configurations are used for the neural networks for each forecast step. Therefore, it is considered important to control the temporal consistency of the combined product. Ideally, spurious jumps are reduced by the combination. However, at least, it should be prevented that additional inconsistencies are introduced by the method of combination.

Several metrics have been introduced to assess temporal forecast inconsistency of a sequence of forecasts. Zsoter et al. (2009) construct a spatial inconsistency index that consists of the differences of two forecast fields normalized by their variability. They then define a "*flip-flop*" as an oscillation of that index of two consecutive forecasts around its mean of the entire sequence of forecasts. The "Forecast Convergence Score" described by Ruth et al. (2009) comprises the count of forecast oscillations around a significance threshold and includes information about the convergence towards the following forecast as well as the magnitude of the oscillations. The "Convergence Index" of Ehret (2010) is a combination of counts of oscillations exceeding a significance threshold and counts of non-convergent forecasts. The metric introduced in Richardson et al. (2020) is based on the average of all ensemble differences of consecutive model initializations. To compute the difference, the divergence function associated with the continuous ranked probability score (CRPS) is utilized. Griffiths et al. (2019) add up the distances between consecutive forecasts over a forecast sequence and divide the sum by the range of the forecasts.

To run and maintain a ML-based precipitation forecasting system in daily operations can be facilitated if the applied architecture of the ML system is simple and robust against changes in the training dataset. Furthermore, the training dataset should contain only few predictors which, besides that, are easy to maintain. Therefore, we would like to address the following issues with the present study. First, we want to assess the forecast quality of the set of hyper-parameter optimized ANNs introduced in Schaumann et al. (2021), when they are trained on an alternative high-resolution dataset. The dataset comprises forecasts of DWD's ensemble-based precipitation nowcasting scheme STEPS-DWD (Reinoso-Rondinel et al. 2022) and ensemble forecasts of an experimental setup of the operational high-resolution short-term NWP model ICON-D2 (ICOsahedron Non-hydrostatic) for the SINFONY-project. Second, we want to explore to which extent the forecast inconsistency (jumpiness) can be reduced by the proposed set of ANN architectures, and whether it is further reduced if only one common ANN architecture is applied to all forecast lead times.

The remainder of the paper is structured as follows. Section 2 gives a brief overview of the utilized datasets. In Section 3, we briefly review some of our previous work and explain which changes are made to the combination model in the present paper. Then, in Section 4, the new combination model is validated and the effects of each change are discussed. Finally, Section 5 summarizes our study and draws some conclusions.

2. Data

The present study assesses the effects on forecast quality, when the set of hyper-parameter optimized ANNs introduced in Schaumann et al. (2021) is trained on a dataset with higher resolution and input forecasts from other ensemble forecast models. For this purpose, we utilize DWD's ensemble-based precipitation nowcasting scheme STEPS-DWD as well as ensemble forecasts of an experimental setup of the operational high-resolution short-term NWP model ICON-D2, both developed in the framework of SINFONY. The training dataset considered in the present study focuses on summertime heavy rainfall events and consists of data for three monthly time periods (from 05/26/2016 to 06/26/2016, from 06/01/2019 to 06/23/2019 and from 06/03/2020 to 07/16/2020). In the following these datasets will be described in more detail.

a. STEPS-DWD

The probabilistic RADVOR (Radarvorhersage; engl. radar forecast) forecasts from our previous study are replaced by the new ensemble precipitation nowcasting method STEPS-DWD. The latter is based on the well-established STEPS approach (Seed 2003; Bowler et al. 2006; Seed et al. 2013; Foresti et al. 2016) and has been adapted and improved for DWD purposes within the framework of SINFONY. The forecasts are based on composites of radar reflectivities obtained by DWD's radar network, which is depicted in Fig. 1 by the envelope of all radar measuring ranges. Furthermore, rain rates are derived by a method for quantitative precipitation estimation (QPE) that uses individual relations between radar reflectivities and rain rates for different types of hydrometeors (Steinert et al. 2021). STEPS-DWD is configured for the present study to consist of a cascade of first-order autoregressive processes on twelve spatial scales and to apply a new localization approach (Pulkkinen et al. 2020; Reinoso-Rondinel et al. 2022) for the estimation of the autoregressive parameters on each individual scale. Individual realizations of the ensemble are then generated by imprinting spatially correlated fields of stochastic noise in regions with precipitation. The spatially recomposed fields are then extrapolated by a constant vector backward scheme (Germann and Zawadzki 2002) based on a predetermined motion vector field. Nowcasts are computed every 30 min out to 6 hours ahead with a temporal resolution of 5 min. The original fields with a spatial resolution of $1 \times 1\,\mathrm{km}^2$ are interpolated onto the coarser NWP grid ($\approx 2.2 \times 2.2\,\mathrm{km}^2$). Afterwards, the extrapolated rain rates are accumulated to hourly rainfall amounts. For lead times less than one hour, accumulations are computed also from radar-based QPE products, so that at the start of each extrapolation forecast, the hourly rainfall amount consists of the radar-based QPE products from the immediately preceding hour. This precipitation accumulation is used as the ground truth.

b. ICON-D2-EPS

Compared to the previous study (Schaumann et al. 2021) in which we used statistically post-processed NWP forecasts as input for the neural network, we now switch to raw NWP ensemble forecasts computed by an experimental setup of the ICON model (Zängl et al. 2015) in limited area mode (LAM) on a central-European domain with a horizontal grid spacing of $\Delta x \approx 2.2\,\mathrm{km}$ and 20 forecast ensemble members. This deep-convection-allowing setup is called ICON-D2. Besides conventional observation data and Mode-S aircraft measurements, 3D volume radar reflectivities and radial winds are assimilated by DWD's kilometre-scale ensemble data assimilation system KENDA, which implements a localized ensemble transform Kalman filter (Schraff et al. 2016; Bick et al. 2016). Note that 40 members are used for the assimilation, while the first 20 members serve as initial conditions for the forecasts. Lateral and upper boundary conditions are provided by ICON-EU ensemble forecasts (larger trans-European domain, grid spacing 6.5 km, parameterized deep convection). For cloud microphysics the operational conventional one-moment scheme is used. Ensemble forecasts are initialized every 3 hours and run up to 12 hours ahead. From these forecasts we use hourly precipitation sums at each forecast hour.

3. Model & Methods

In Schaumann et al. (2021), an ANN-based model for the combination of two probabilistic forecasts, which produces calibrated and consistent probabilities, has been proposed as a generalization of the so-called LTI-model (Logistic regression, Triangular functions, and Interaction terms; see Schaumann et al. (2020)). The notion "probabilistic forecast" refers to probabilities for the occurrence of binary events, i.e., the exceedance of precipitation thresholds. With

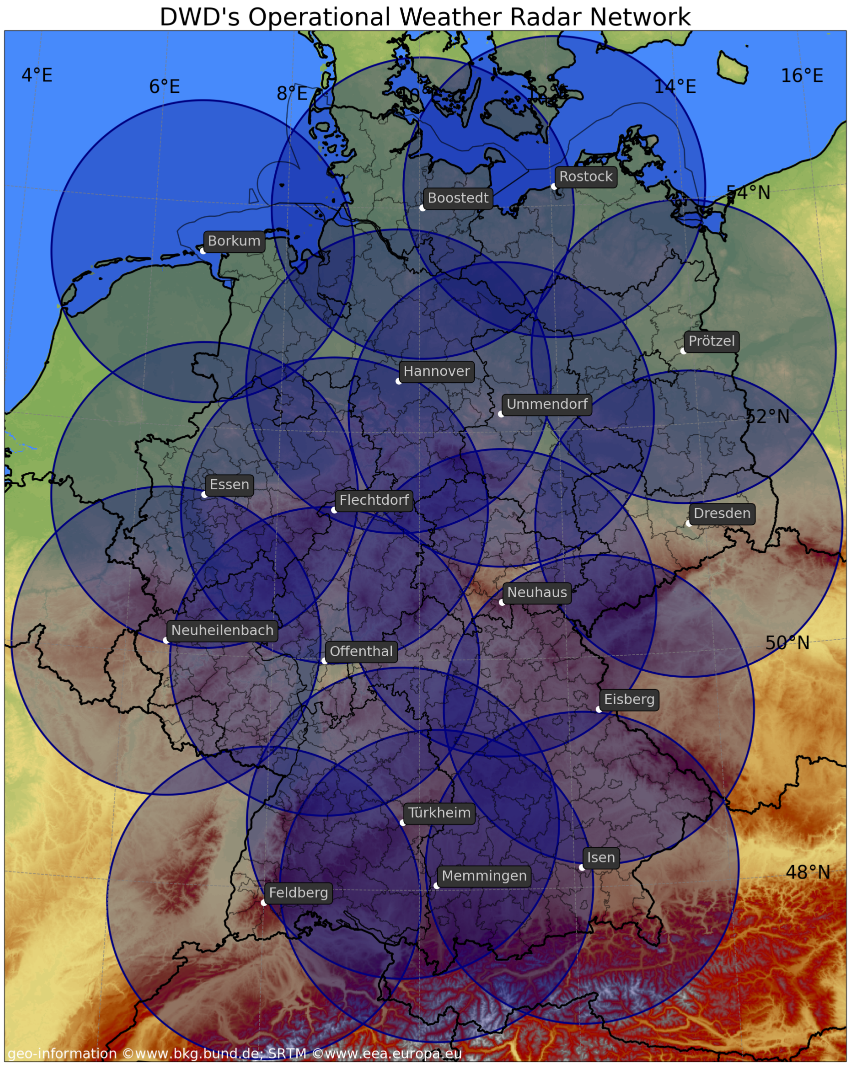


FIG. 1: DWD's operational radar network. The positions of radar sites of the network are depicted by white dots. The blueish circle around each site represents the range of the terrain-following precipitation scan of the dual-polarization radars. Darker blue shades represent areas covered by more than one radar. Additionally, the terrain height is illustrated in colors from green (low) to white (high). Note that the radar at Borkum, at the time of the training period, has been an older single-polarization radar with a lower data quality.

a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a probabilistic forecast $P : \Omega \longrightarrow [0,1]$ for a random event $E : \Omega \longrightarrow \{0,1\}$ is considered to be calibrated when $p \approx \mathbb{E}(E|P = p)$ for all $p \in [0,1]$. This is a desirable property since, colloquially said, we expect the event to occur with the forecasted probability. The relationship between p and $\mathbb{E}(E|P = p)$ of a forecast model is expressed in its reliability diagram. When considering a set of probabilistic forecasts $P_1, P_2, \ldots$ for binary events

$E_1, E_2, \ldots$ with $\mathbb{P}(E_1 \geq E_2 \geq \ldots) = 1$, we call the forecasts $P_1, P_2, \ldots$ consistent if it holds that $\mathbb{P}(P_1 \geq P_2 \geq \ldots) = 1$. In case of the exceedance of increasing precipitation thresholds, we know that a higher threshold can only be exceeded, if all lower thresholds are exceeded, too, and therefore the forecasted probabilities should be monotonously decreasing.

In this section, we propose a few improvements of the ANN-model and call the new version C^3-model (where C^3 means combined, calibrated, consistent).

a. Architecture and Properties of the C^3-Model

In its current form, the C^3-model consists of several feed-forward neural networks, each one for the combination of forecasts with respect to a specific lead time.

All neural networks considered in the C^3-model consist of 4 types of layers arranged in the following order: 0 to 5 convolutional layers, one dense layer, one triangular functions layer and one dense layer with softmax activation function, see Fig. 2.

The triangular functions layer transforms each scalar x_i of its input in such a way that each neuron j of the following dense layer can be interpreted as a sum of functions $S_{j,1}(x_1) + \ldots + S_{j,n}(x_n)$ where the functions $S_{j,i} : [0,1] \to \mathbb{R}$ are linear splines determined by the weights of the dense layer. Compared to the the linear combination of inputs $w_1 x_1 + \ldots + w_n x_n$ without triangular functions, a sum of linear splines gives the network far more flexibility in how it models the relationship between the inputs $x_1, \ldots, x_n$ and the outputs of the following dense layer.

As a loss function the categorical cross-entropy is used. The softmax layer produces a discrete probability distribution for the events that precipitation occurs either between two consecutive thresholds or below/above the lowest/highest threshold. Based on the discrete probability distribution a probability for the exceedance of each threshold is computed. The combined forecast is calibrated and consists of consistent probabilities for each precipitation threshold.

The specific hyper-parameters of each neural network are determined individually by a hyper-parameter optimization algorithm, see Table 1. For more details about the triangular layer or the hyper-parameter optimization algorithm, see Schaumann et al. (2021).

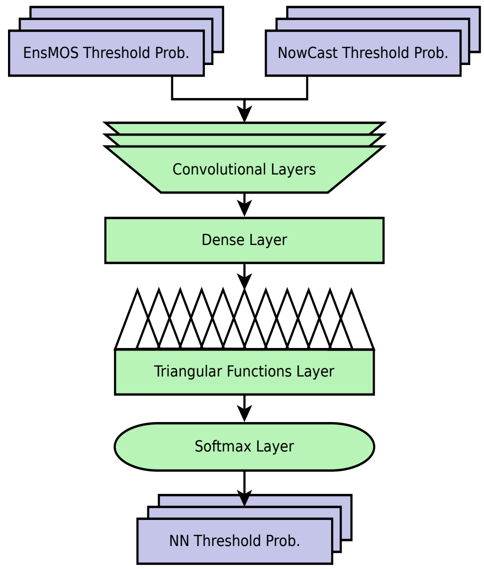


FIG. 2: The utilized network architecture (green) and the input and output data (blue) as a schematic representation adapted from Schaumann et al. (2021). The arrows depict the flow of information. Note that the input data in this study are forecasts from STEPS-DWD and ICON-D2.

b. Training and Validation

The first 3 weeks of the available dataset are used as a warm-up period on which the model is trained only. For the remaining dataset, a rolling-origin scheme (Armstrong and Grohman 1972) is applied. This is an iterative approach that simulates how new information becomes incrementally available for training over time in an operational setting. The available dataset is divided by a point in time t into two parts, where t represents the presence. The part of the dataset before t is considered to be in the past and therefore available for training, while the data after t is considered to be future and is used for the validation of the model. Over the course of the training and validation process, t is iteratively moved from the beginning to the end of the dataset, where in each step of the rolling-origin scheme the model is trained on the past data and validated on the future data.

c. Increase in Spatial Resolution

In the present paper, in order to increase the resolution of the combined forecast, two new initial forecasts are considered: STEPS-DWD and ICON-D2. Both have a spatial resolution of

Lead time	No. of conv. layers	Kernel size	Conv. activation	Conv. reg.	Conv. output length
+1h	4	3	elu	5e-07	14
+2h	4	3	elu	5e-06	6
+3h	4	3	relu	0	8
+4h	1	9	elu	0	4
+5h	1	9	elu	5e-06	12
+6h	2	6	sigmoid	0	4

Lead time	No. of neurons in dense layer	Dense activation	Dense reg.	No. of triangular func.	Optimizer
+1h	12	sigmoid	0	9	Nadam
+2h	12	tanh	0	5	Nadam
+3h	12	exponential	1e-06	3	Adamax
+4h	6	tanh	0	5	Adamax
+5h	10	sigmoid	0	3	Adamax
+6h	10	relu	1e-05	5	Adamax

TABLE 1: Selected configurations of hyper-parameters for different lead times based on the results of Schaumann et al. (2021).

$2.2 \times 2.2\,\mathrm{km}^2$, whereas the datasets previously used in Schaumann et al. (2021) have a resolution of $20 \times 20\,\mathrm{km}^2$. As the input of the ANN consists of data for a fixed number of grid points determined by the convolutional layers, the spatial range of the input data shrinks when the resolution of the datasets is increased. To compensate for the finer grid, one could increase the size or dilution of the convolutional layers. Here, dilution refers to a method, where only every N-th row and column of the input data is passed on to the network, i.e., a dilution of 2 doubles the spatial range of convolutional layers along the x- and y-axis without increasing the number of data points. However, it turned out that adapting the size of the convolutions to the finer grid did not result in better validation scores. Similarly, diluting the convolutions did not lead to better validation scores and additionally introduced artifacts to the combined forecast. Due to the gaps introduced by a dilution of $N > 0$, smaller structures in the input forecast influence every N-th grid point in the output, while neighbouring grid points are unaffected. This leads to repeating patterns in the combined forecast. Therefore we did not change the convolutions for the results obtained in the present paper.

d. Full Utilization of the Sampling Window with an NaN-Mask

Due to technical reasons, the neural network in Schaumann et al. (2021) requires its input to have a rectangular shape and no missing data. Since the precipitation nowcasting forecasts are based on radar composites, which are shifted according to a motion vector field, parts may be shifted outside of the sampling window. Temporary radar outages further reduce the available data. Depending on shape and location of the area with available data, the largest usable rectangle might be considerably smaller than the area with available data itself. To utilize the whole dataset for the present paper, all NaNs are replaced by the value -1 and a Boolean field which flags grid points with missing data is used as an additional input to the neural network. Instead of discarding part of the dataset, this approach allows the model to learn to ignore the values of -1.

e. Forecast Persistence and Consistency

Repeated runs of a forecast model at different starting times produce a sequence of forecasts with different lead times for the same valid date. In general, these forecasts become increasingly accurate with decreasing lead time and the ideal evolution would be a trend from inaccurate or climatological values to a more accurate forecast with decreasing lead times. However, due to random (non-systematic) forecast errors, the trend is often not monotonous for the individual cases. Sometimes, older forecasts are more accurate than newer updates and spurious jumps in the forecasts appear. These inconsistencies or *jumps* are especially harmful for warning management. A weather warning that is issued for a specific date and time, canceled later on (based on a new forecast run), and then possibly issued again with the next forecast, is not considered trustworthy and can hardly be communicated to the public.

In the present study, we consider the flip-flop index FFI introduced by Griffiths et al. (2019) as a metric for temporal forecast consistency and investigate how the input forecasts and the combined product behave with decreasing lead time. This index is defined by

$$\mathrm{FFI}(V_{i,j}) = \frac{\sum_{l\in\{1,\ldots,L-1\}} |v_{l+1} - v_l| - (\max_l(V_{i,j}) - \min_l(V_{i,j}))}{L-2}, \tag{1}$$

where $V_{i,j} = (v_1, \ldots, v_L) \in \mathbb{R}^L$ is a vector of predictions for grid point i, j and for the same valid time, with $L > 2$ lead times. The FFI is normalized by $L-2$, which is the length of the vector minus first and last point and thus the maximum number of predictions that can be not monotonous with respect to their respective predecessor and successor. Note that $\mathrm{FFI}(V_{i,j}) = 0$ indicates a perfect flip-flop index and is achieved by forecasts that converge monotonously in time. Forecasts with an oscillating pattern V are penalized resulting in a flip-flop index $\mathrm{FFI}(V_{i,j}) > 0$.

The FFI is evaluated for each grid point i, j individually and averaged over the whole evaluation period as mentioned in Section 4c .

4. Results

a. Lead-Time Dependent Investigation on Model Performance

We want to assess whether the C^3-model with its new implementations and the high-resolution input datasets is still able to produce high-quality forecasts, where we want to emphasize the core features of its forecasts: combination, calibration, and consistency. For this, we computed the bias, Brier skill score, reliability, sharpness, and the area under ROC (relative operating characteristic) curve (AUC) over the whole evaluation period for the C^3-model with two different hyper-parameter settings to assess the importance of lead time dependent hyper-parameters (C^3: lead-time dependent, and C^3_{LT1}: only from lead time +1 h) as well as for both individual input forecast systems STEPS-DWD and ICON-D2. We chose these metrics to get some easily interpretable indicators regarding the

1. systematic model error
2. forecast quality in terms of systematic model errors and random forecast errors
3. conditional frequency bias
4. forecast resolution
5. discrimination ability

If we call the exceedance of an arbitrary threshold an event, all of these metrics are based on the grid-box-wise actually observed event occurrence and/or the forecasted event probability. Thus, the bias describes the mean error (ME) of the forecasted event probabilities and indicates the unconditional systematic model error. The Brier skill score (BSS) consists of the mean squared error (MSE), representing the Brier score (BS) itself, divided by a reference BS that is based on the sample climatology. The values of BSS reach from $-\infty$ up to 1, whereas values of 1 and 0 depict a perfect forecast and the climatologic forecast, respectively. Further, the reliability indicates how far the reliability diagram of a forecast deviates from the ideal line, i.e., the reliability is the weighted mean of the squared differences between the reliability diagram and the ideal line for each bin, where the weights are the number of forecasts within each bin. Ideally, the predicted probability is equal to the observed relative frequency in which case the reliability diagram is equal to the ideal line and the reliability is equal to zero. Therefore, the reliability provides information about the frequency bias of the forecasted event probabilities and represents a measure for the calibration of an ensemble forecast. The sharpness characterizes the unconditional distribution of the probability forecasts and provides information about the forecast resolution, i.e. the ability to predict extreme values close to 0 or 1. It is represented by the variance of the forecasts. The last metric considered is the AUC, which provides information about a forecast's ability to discriminate between events and non-events.

The results obtained for the bias are depicted in the first column of Fig. 3 for both configurations of the C^3-model, STEPS-DWD, and ICON-D2, as a green, red, yellow, and blue line, respectively. All four forecast techniques exhibit a nearly lead-time independent systematic error. However, the event probability for the lowest threshold of 0.1 mm hourly rainfall amount is overestimated by ICON-D2 forecasts by 1 percentage point, whereas the extrapolations of STEPS-DWD reveal a slight underestimation by 1 percentage point. For higher thresholds these systematic errors of STEPS-DWD and ICON-D2 diminish as the frequency of event occurrences within the evaluation period decreases. The forecasts of both configurations of C^3 are bias-free in the first two hours. Afterwards, they exhibit only a small difference from zero, whereas C^3 tends towards ICON-D2, and C^3_{LT1} towards STEPS-DWD. These results imply that the model is able to reduce systematic errors caused by the input forecasts with respect to the ground truth. This bias correction can be seen as one part of a forecast calibration.

To assess the forecast quality in due consideration of the combination aspect, the results obtained for the BSS are illustrated in the second column of Fig. 3, in the same way as the bias. Here, the BSS of the precipitation nowcast extrapolations of STEPS-DWD starts with a high skill, since they start from the observation, but decrease rapidly since growth and decay processes of precipitation are not represented. Errors in initial and boundary conditions cause that the NWP forecasts of

ICON-D2 start with a lower skill. However, the decrease with increasing lead time is not that pronounced due to the explicit simulation of the dynamical evolution. The intersection of both curves denotes the point when the quality of nowcast extrapolations sinks below those of the NWP forecasts and occurs around 2.5 h after initialization. The forecast quality in terms of BSS of both C^3-models outperforms each individual input forecast technique at all lead times. Furthermore, the different hyper-parameter settings only have a small effect on forecast quality, so that an optimal combination of the two input forecasts is achieved with both C^3-models at all lead times. However, the approximately 0.1-higher BSS values should be treated with caution, since convolution-induced spatial smoothing of the forecasts (as discussed by, e.g. Cuomo and Chandrasekar (2021)) leads to better scores of continuous verification metrics.

If such smoothing effects decisively affect the forecasts of our C^3-models, it should be visible in the reliability, since the frequency of predicted high probabilities is decreased, whereas the frequency of observed events for forecasted intermediate probabilities increases. At first, the area enclosed by the reliability curves and the aforementioned ideal course is depicted for each of the forecast systems in the third column of Fig. 3, again in the same manner as the bias. For STEPS-DWD, this area size reveals an increase with lead time, whereas that of the ICON-D2 forecasts is nearly constant. Note that we utilize raw NWP output data that is uncalibrated. The area size of both C^3-models is smaller than those of the two input forecasts indicating that the curves are closer to the ideal course and, therefore, the combined forecasts are more reliable than the forecasts of the input systems. However, the size of the area fluctuates for the two C^3-models at later lead times. This may be an indicator for shortcomings in the calibration due to the choice of triangular functions. The forecast calibration depends on the number of these triangular functions, which is equal to 9 for +1 h and only 5 and 3 for +4 h/+6 h and +5 h, respectively (cf. Tab. 1).

Nevertheless, the forecast sharpness of both C^3-models is reduced compared to the forecasts of ICON and STEPS-DWD for all exceedance thresholds and lead times, as shown in the fourth column of Fig. 3. This may have several reasons. On one hand, the raw input forecast ensembles reveal a wider range of probabilities even for hourly rainfall amounts above 5 mm. Thus, the sharpness is increased, but at the expense of reliability. On the other hand, the increasing forecast uncertainty for higher lead times and the training on less frequent events lead to a loss of probabilities close to 1, which reduces the sharpness. However, the C^3-models exhibit a higher AUC compared to ICON and STEPS-DWD forecasts, which is shown in the fifth column of Fig. 3. This may indicate an improved discriminating ability between events and non-events, albeit this result should be treated with caution. Not only the missing high probability values, but also the low event base rate may be misleading. A further investigation on the discrimination ability of the C^3 forecasts based on an improved AUC as described by Ben Bouallègue and Richardson (2022) may provide more reliable results.

To get a more detailed insight at the conditional bias, the reliability diagrams of the four forecast systems are depicted in Fig. 4 for the lead times +1 h, +3 h, and +6 h and five thresholds from 0.1 mm up to 5 mm. In addition, below each reliability diagram, the frequency histograms for each of the forecasts are depicted to give an evaluation of the forecast sharpness. Both the extrapolation nowcasts of STEPS-DWD and the forecasts of ICON-D2 are overconfident over the entire range of thresholds and lead times. The overconfidence of ICON-D2 forecasts increases especially with the threshold since the frequency of observed events is not only lower due to the higher threshold but may also be reduced due to errors in location. Besides the missing representation of the dynamics of precipitation in STEPS-DWD forecasts, their spread is small leading to that overconfidence. The results obtained for both combination models are well calibrated for all depicted thresholds at a lead time of +1 h. With increasing lead time, the C^3_{LT1} forecasts remain calibrated, though, high probabilities are no longer forecasted. However, at a lead time of +3 h, the forecasts of the C^3-model exhibit structures that can be connected with the spatial smoothing. For the C^3_{LT1} model, such a structure is only visible for a threshold of 2 mm at a lead time of +6 h.

These results show that even with the new dataset forecasts of both models C^3 and C^3_{LT1} are well-calibrated in terms of bias correction and reliability, consistent for the range of thresholds, and that they are composed of the optimal combination of the two input forecast systems STEPS-DWD and ICON-D2. Furthermore, the forecast calibration may be able to reduce the impact of convolution-induced spatial smoothing.

b. Investigation of Spatial Patterns in Systematic Model Error and Forecast Quality

We want to investigate possible reasons for systematic model errors in the forecasts of STEPS-DWD and ICON-D2 and how the combined forecasts reduce these errors. Further, we want to explore whether spatial patterns are visible in the forecast quality. For this, the spatially resolved

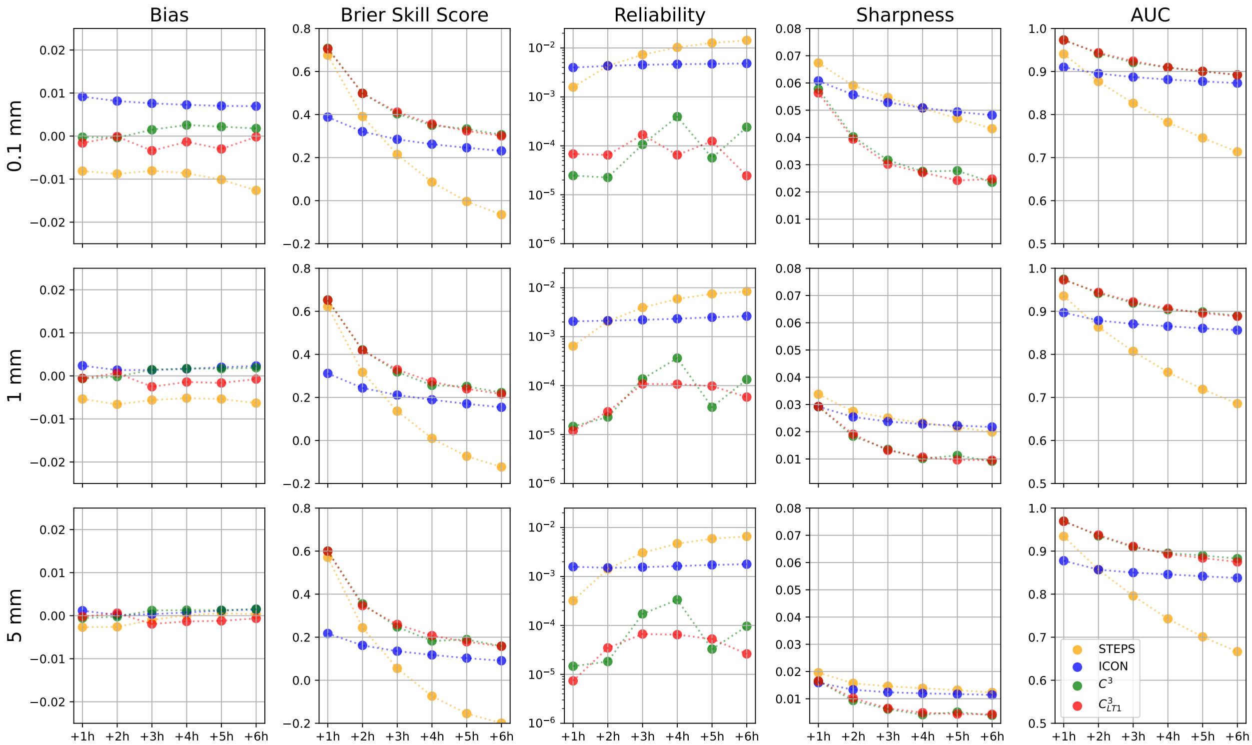


FIG. 3: Bias (in %) (first column), Brier skill score (second column), reliability (third column), sharpness (fourth column), and area under the ROC (relative operating characteristic) curve (AUC; fifth column) averaged over the evaluation period as validation scores for three of the nine considered thresholds for the combfination model with lead-time-dependent hyper-parameters (C^3; green), the combination model with hyper-parameters for a lead time of +1 h (C^3_{LT1}; red), STEPS-DWD (STEPS; yellow), and ICON-D2 (ICON; blue).

biases for lead times of +1 h and +3 h are illustrated in Figs. 5a and 5b, respectively. The spatially resolved BSS is depicted in Figs. 6a and 6b for identical lead times.

The results for STEPS-DWD are shown in the left columns of Figures 5 and 6. The nowcast extrapolations exhibit at a lead time of +1 h that the slight underestimation discussed in the previous section occurs almost over the entire domain. Only in regions close to radar sites which are covered by a single radar (cf. Fig. 1) an overestimation is visible. The underestimation may be caused by a loss of power induced by the way the spatially correlated stochastic noise fields are generated, see e.g. Atencia and Zawadzki (2014). The overestimation may be due to rain rates estimated in different heights, e.g., when rain rates estimated at maximum range of a given site are advected and compared to near-surface estimates of the respective radar site. A further reason could be attenuation caused by heavy precipitation directly at the radar site. With increasing lead time, errors due to the forecast field shifting lead to an underestimation in the western and southern

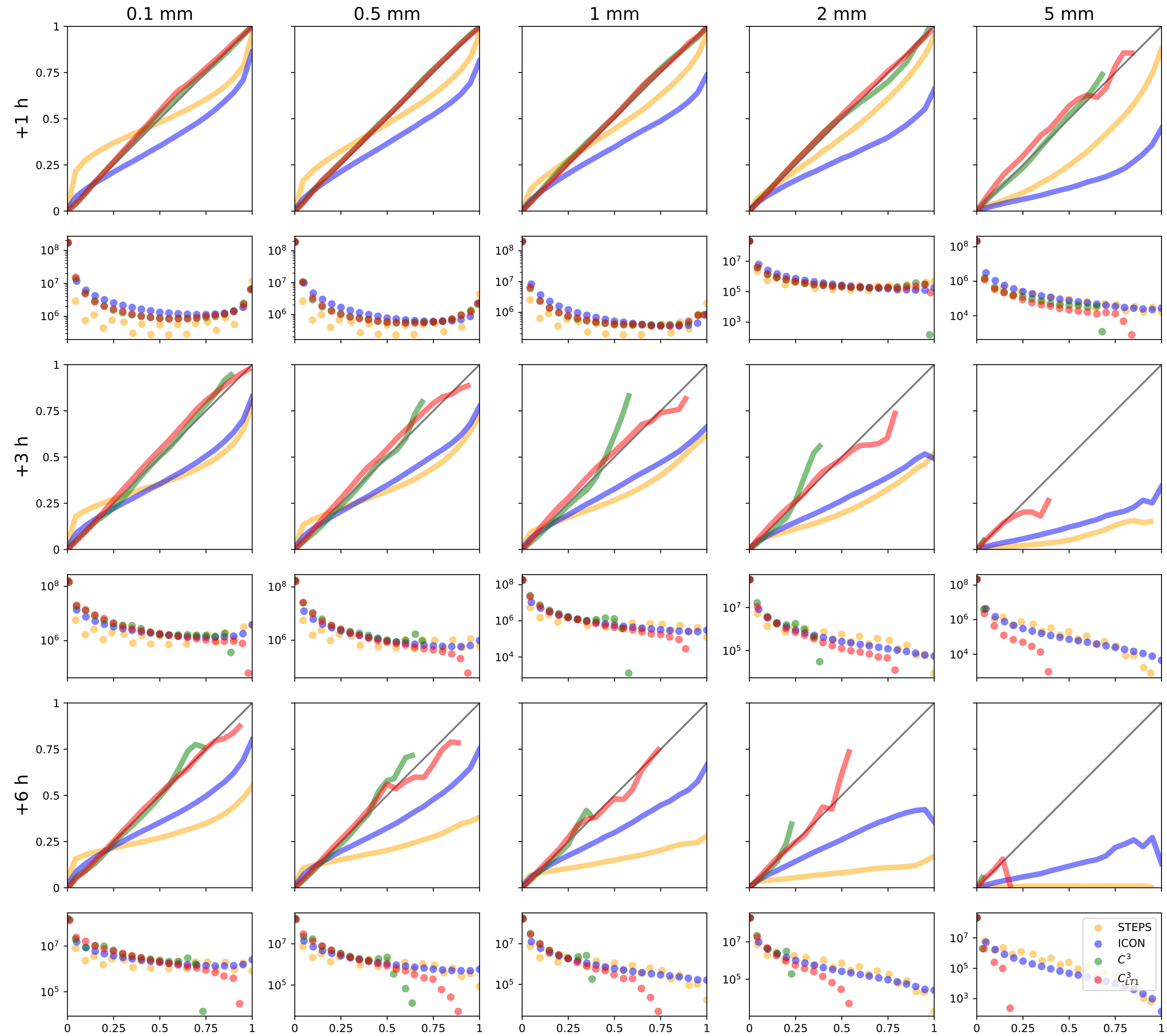
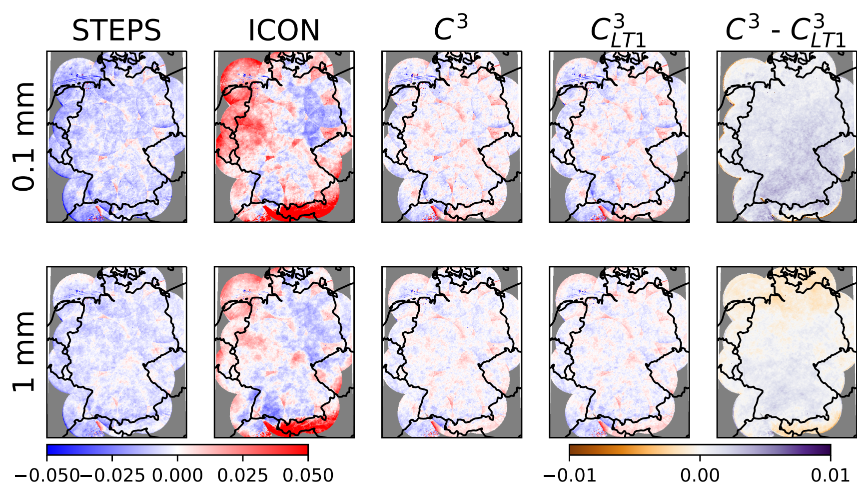


FIG. 4: Reliability diagrams for the C^3-model with different architectures (C^3; green), the C^3-model with one architecture (C$^3_{\mathrm{LT1}}$; red), STEPS-DWD (STEPS; yellow), and ICON-D2 (ICON; blue). Here, the x-axes indicate the forecast probability and the y-axes indicate the observed frequency of the event. Below each reliability diagram, the frequency histograms for each of the forecasts are depicted.

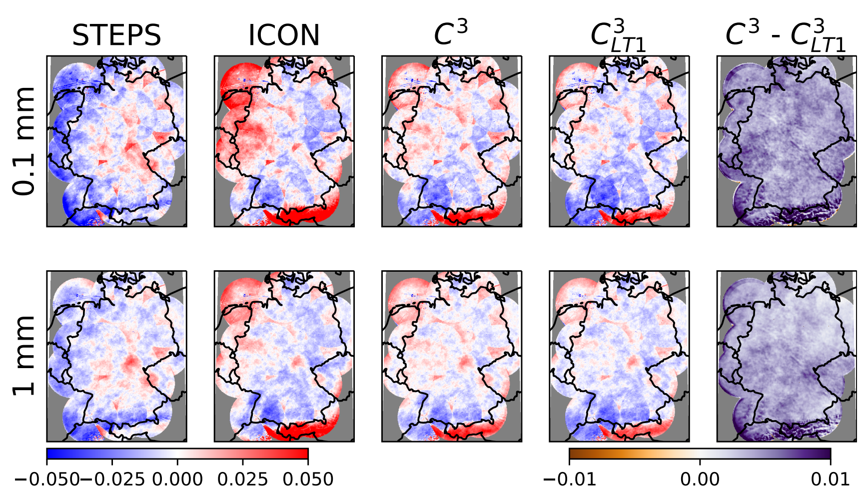
part of the domain, whereas the missing of dynamical evolution increases the systematic error in the entire domain. The BSS reveals no distinct spatial pattern. However, some of the radar sites and also the Alps are visible, which is caused by a lower event occurrence due to radar outages, attenuation effects and/or beam blocking. With increasing lead time, the aforementioned strong decrease in the BSS is visible.

The spatially resolved results for ICON-D2 forecasts are shown in the second columns from the left of Figures 5 and 6. Here, the bias not only depicts systematic model errors but also systematic errors between the simulated surface precipitation sum and the QPE used as ground truth. Therefore, the overestimation mentioned above can be attributed to typical radar and compositing shortcomings. First, the strong overestimation over the Alps is caused by beam blocking. Second, range attenuation and ground clutter can be seen at Borkum radar site in the north-west due to a positive bias at long ranges and a local negative bias close to the radar site. For the radar site located at the Feldberg in the south-west a height difference where rain rates are simulated/estimated may be a reason for the underestimation of hourly precipitation sums. This underestimation is more distinct for a threshold of 1 mm. Nevertheless, the overestimation in the western part of the domain, which is noticeable especially for the threshold of 0.1 mm, could be attributed to meteorological phenomena. For a higher lead time the main patterns remain the same, however, with a higher magnitude. In addition to the aspects mentioned above for the spatially resolved BSS of STEPS-DWD, the range attenuation is more distinct for the BSS of ICON-D2, especially at a threshold of 1 mm. For later lead times, the decrease is not that pronounced as for STEPS-DWD.

The spatially resolved results for the C^3-models are shown in the right columns of Figures 5 and 6. One the one hand, it can be seen that the biases are reduced in terms of magnitude compared to both input forecast systems for +1 h lead time and both thresholds of 0.1 mm and 1 mm. However, the spatial patterns induced by the shortcomings of the radar-based QPE composite are still visible in the bias. On the other hand, even at this lead time some spatial patterns are comparable to those of ICON-D2. For example, the overestimation induced by beam blocking at the Alps and the range attenuation of the Borkum radar. This means that the deficits of the composite used as ground truth, e.g., lower estimated rainfall amounts in regions covered just due to one radar and deviations due to ground clutter, are not learned by the C^3-model. In addition, the C^3-model forecasts are constrained by both input forecasts so that they are the result of the best possible combination. For a lead time of +3 h, the spatial bias patterns are closer to that of ICON-D2-EPS, although the overestimation in the western part of the domain is reduced and the underestimation in the range of the Feldberg radar is even more pronounced for the threshold of 0.1 mm. Therefore, even with a higher weighting towards ICON-D2-EPS, the systematic overestimation in the western part is reduced by the C^3-model. However, the differences in hourly rainfall amount caused by the aforementioned height difference of the Feldberg radar is not reduced since STEPS-DWD forecasts exhibit an underestimation as well. The systematic error of the C$^3_{\mathrm{LT1}}$-model is only slightly below that of the C^3-model, which is shown in the right column of Fig. 5. The spatially resolved values of BSS of the C^3-forecasts are above the BSS of each input forecast systems for both thresholds and lead times in the entire domain. The spatial structures caused by the QPE composite are also visible in the results of the C^3-model, additionally indicating that those shortcomings are not learned by the C^3-model.

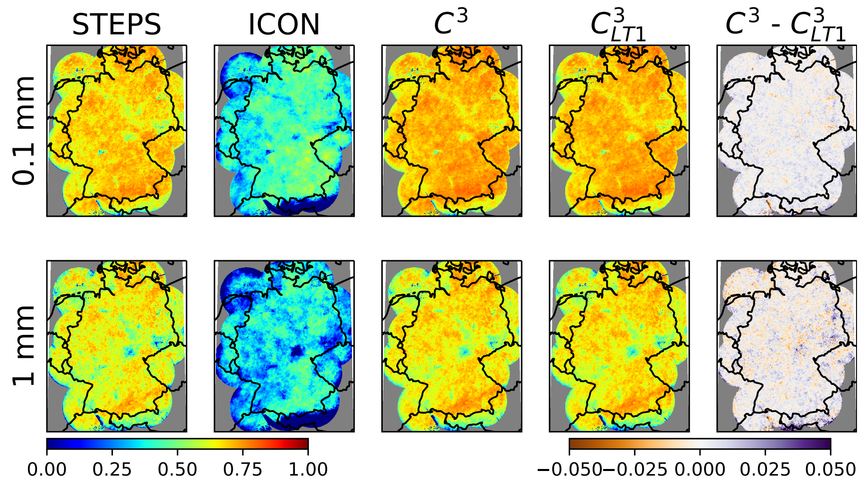


(a) Lead time: +1 h

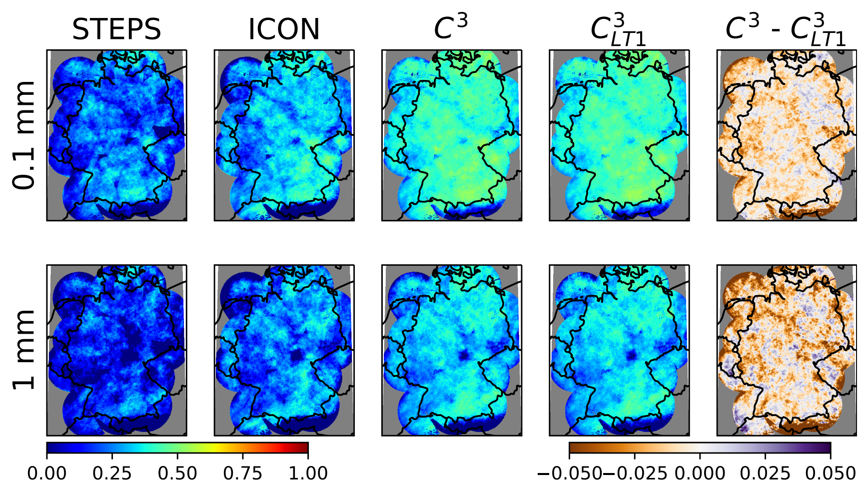


(b) Lead time: +3 h

FIG. 5: Spatial distribution of the bias (in %) averaged over the considered period for (a) +1 h, and (b) +3 h lead time. Depicted are STEPS-DWD (left column), ICON-D2 (center column), and the combination model (C^3) with different architectures (center right columns). The right column shows the difference between C^3 and C$^3_{\mathrm{LT1}}$.



(a) Lead time: +1 h



(b) Lead time: +3 h

Fig. 6: The same as Fig. 5 but for the Brier skill score.

c. Temporal Consistency of the Combined Forecasts

Another question is how the temporal consistency of the combined forecasts compares to those of both initial forecasts, and how it is affected by the hyper-parameter choice of the C^3-model. The flip-flop index (FFI) given in Eq. (1) is averaged over the evaluation period and visualized in Fig. 7 for both initial forecast systems STEPS-DWD (STEPS) and ICON-D2 (ICON), the combination model (C^3), as well as the modified combination model (C^3_{LT1}). The C^3-model has lead-time dependent hyper-parameters in order to provide maximal adaptation. To control whether lead-time dependent architectures affect temporal consistency for a sequence of forecast updates that are valid for the same date, the C^3_{LT1}-model uses the hyper-parameters of the lead time +1 h of the C^3-model for all lead times +1 h, . . . , +6 h.

The FFI of the probabilities for the events that hourly precipitation exceeds the thresholds 0.1 mm and 1 mm is presented in Fig. 7a as average over the evaluation period. To better understand what the FFI values mean in our case, a brief example is given. A sequence of forecasts is optimal in terms of temporal consistency if it follows the shortest distance between its minimum and its maximum. This distance is scaled by the maximum number of possible flip-flops within the sequence and is used as a reference value. A FFI of 0.03 indicates that the difference in event probabilities between two consecutive forecasts at a given grid box is on average 3 percentage points larger than the reference value. Many cases with no precipitation in forecast and observation reduce the average FFI. To account for this effect, Fig. 7b depicts the average FFI under the condition that the observed hourly precipitation is at least 0.1 mm, which leads to much larger values compared to the unconditional FFI in Fig. 7b.

The technique of STEPS-DWD consists of two main components which may affect the temporal consistency in different ways. First, a set of first-order autoregressive processes is considered which replace signals on spatial scales that are no longer predictable by spatially correlated noise. Second, an advection scheme is used which extrapolates the forecast fields based on a predetermined motion vector field. As can be observed in Fig. 4, STEPS-DWD forecasts are overconfident especially for longer lead times and higher thresholds, a convergence from climatological event probabilities towards observed event frequencies seems to appear less likely. Moreover, differences between estimated motion vector fields (e.g., lower magnitude, errors in direction) for different lead times may lead to spatial shifts of the predicted precipitation pattern. These spatial shifts may lead to a double penalty problem, when the precipitation patterns of two consecutive forecasts do not align, i.e., both predict precipitation at two different locations, which results in high absolute differences between both locations. Additionally, the temporal evolution of precipitation is not covered by such an extrapolation forecast, i.e., the observed stage of precipitation is extrapolated in time ignoring growth and decay processes. Therefore, any difference in observed precipitation frequency between two consecutive hours leads to an increase in FFI. Furthermore, due to the most common westerly winds and the advection of precipitation, values at the west border of the domain fade out with constant advection, since the precipitation data covers only Germany. At a threshold of 1 mm, effects of beam blocking and range attenuation, as discussed above for bias and BSS, are more apparent.

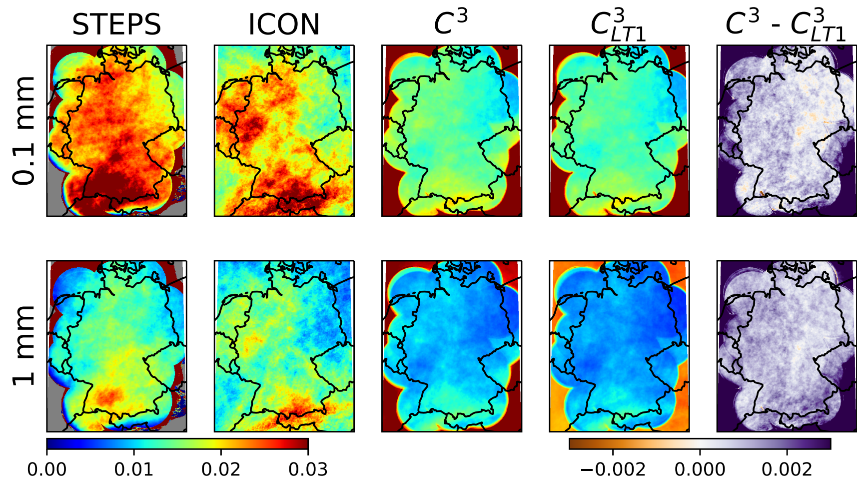
The temporal consistency of NWP precipitation forecasts in convective situations may be affected by the time of convective initiation, the simulated dynamical evolution of precipitation, and also by the location of airmass boundaries or convergence lines. When we consider the unconditional FFI for the ICON-D2 (ICON) forecasts in Fig. 7a, they reveal two regions in which the unconditional FFI is elevated. First, this is a region in Northwestern Germany which can be attributed to uncertainties in the location of airmass boundaries or convergence lines. This can also be seen in the conditional FFI in Fig. 7b. Second, this concerns the upland regions and the Alps which could be an indicator for the prediction uncertainty of orographically induced precipitation. However, this is less pronounced in the conditional FFI, where the largest values can be found in Bavaria. One reason for the reduction of the conditional FFI compared to the unconditional FFI over the Alps may be the frequency bias in observed events due to the previously discussed beam blocking.

Both combination models significantly improve the FFI for both thresholds and with respect to unconditional and conditional averaging. The C^3-model has slightly higher FFI values than the C^3_{LT1}-model, reflecting that the lead-time dependent architecture contributes to the FFI. Compared to the much larger FFI values of both input systems, however, this contribution can be assumed to be insignificant. Moreover, temporary radar outages affect the unconditional FFI, which can be seen in the eastern part of the domain for the radar sites Dresden and Eisberg, cf. Fig. 1.

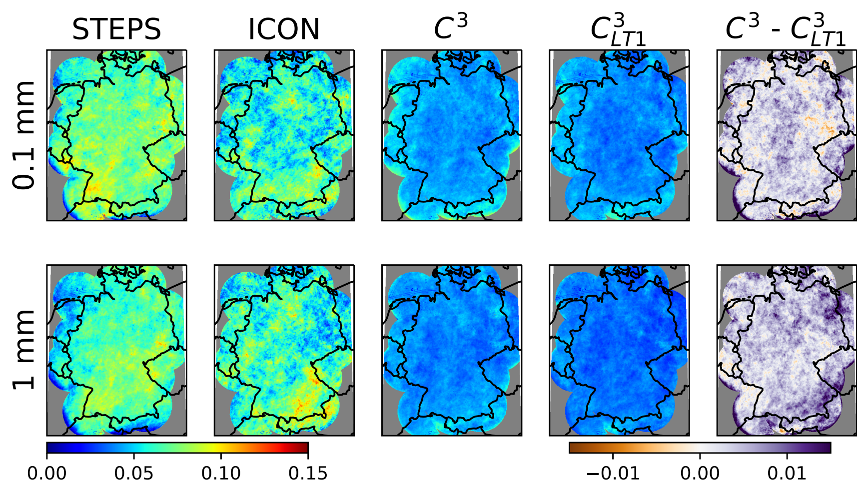
d. Forecast Animations

The supplementary material includes animations showing +3 h probabilistic forecasts of the input techniques STEPS-DWD (STEPS; second column from the left) and ICON-D2 (ICON; third column) as well as the combination model (C^3; fourth column) and the modified combination model (C^3_{LT1}; fifth column). Fig. 8 illustrates an example of these animations. Depicted are the exceedance probabilities for hourly precipitation of at least 0.1 mm (upper row) and 1 mm (lower row). The forecasts were initiated at 1400 UTC of June 4, 2020, and the corresponding observed threshold exceedances are shown in the first column of Fig. 8.

The STEPS-DWD forecasts exhibit less spread at a threshold of 0.1 mm compared to those of ICON-D2. This corresponds to the results shown in the reliability diagrams of Fig. 4, where



(a) Average flip flop index



(b) Average flip flop index under the condition that precipitation is above 0.1mm

FIG. 7: Flip-flop index (FFI) averaged (a) over the whole evaluation period and (b) over the evaluation period under the condition that the observed hourly precipitation is at least 0.1 mm. Note that the color scales of each subplot cover different value ranges. The FFI is depicted for both initial forecast systems STEPS-DWD (STEPS) and ICON-D2 (ICON), the combination model (C^3) as well as the modified combination model (C^3_{LT1}) using the hyper-parameters determined for +1 h for all lead times. The right column shows the difference between C^3 and C^3_{LT1}. For the sake of clarity, only the flip-flop indices for the thresholds of 0.1 mm and 1 mm are shown.

STEPS-DWD is overconfident. However, the area covered by probabilities is close to that of the observation, showing that the dynamical evolution of the precipitation field in this case barely affects the extrapolation forecast. Solely, the precipitation band from Switzerland to Bavaria is less covered by STEPS-DWD. In contrast, this precipitation band is more pronounced in the ICON-D2

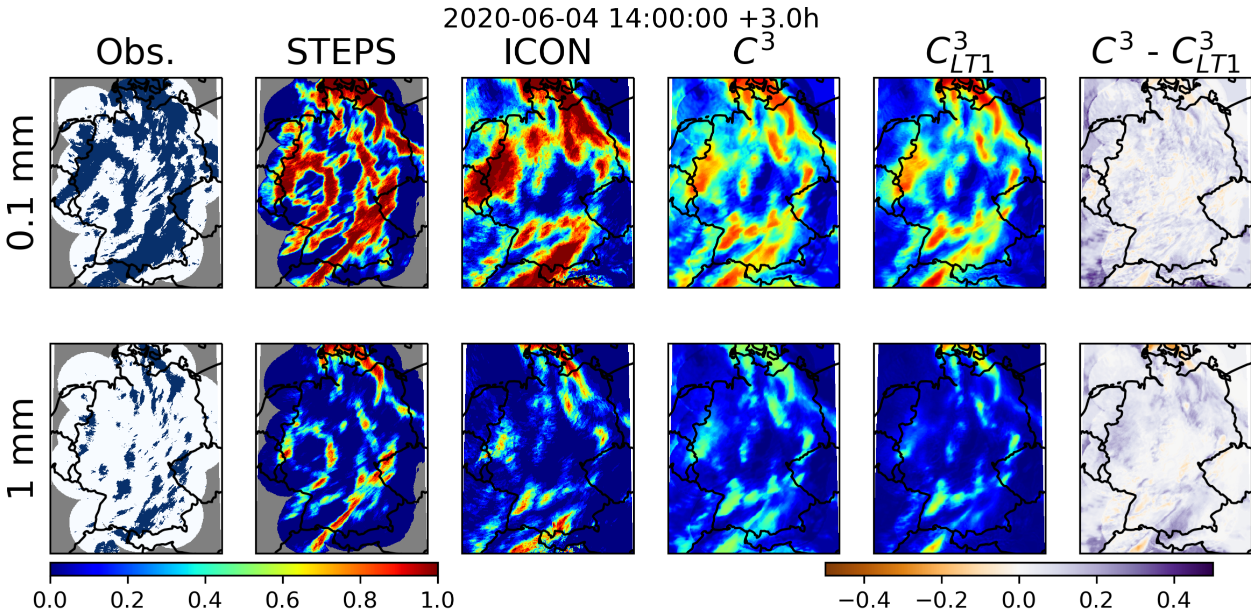


FIG. 8: Exemplary +3 h probabilistic forecasts of STEPS-DWD (STEPS; second column), ICON-D2 (ICON; third column), the combination model (C^3; fourth column), and the modified combination model (C^3_{LT1}; fifth column) for hourly precipitation thresholds of 0.1 mm (upper row) and 1 mm (lower row) initiated at 1400 UTC of June 4, 2020. The corresponding observed threshold exceedances are depicted in blue in the left column. The right column shows the differences between C^3 and C^3_{LT1}. Missing data is marked grey.

forecast. However, the observed precipitation in the center of Germany is not predicted by the ICON-D2 forecast.

Both combination models, C^3 and C^3_{LT1}, exhibit a robust mixture of both input forecasts and provide rather similar results for a threshold of 0.1 mm. Artifacts at the edges of the radar network and also small-scale NWP features (e.g. over the Vosges) are more pronounced in the C^3-forecast. However, the probabilities of the C^3_{LT1}-forecast are higher for both thresholds. Especially for the 1 mm threshold one can see a maximum of about 0.5 for the C^3-model. This corresponds to the results shown in the reliability diagrams of Fig. 4 and can be attributed to the low number of triangular functions, cf. Tab. 1.

5. Conclusions

a. Summary of results

Considering forecasts of hourly rainfall of an advection-based precipitation nowcasting ensemble and of a NWP ensemble system, one can have on the one hand radar outages or relocations of radar sites and on the other hand updates of the NWP model. A simple architecture in combination with a rolling-origin training scheme can made a ML-based seamless precipitation forecasting system robust against those changes in the training dataset and is thus able to support the operational running of a forecasting system. In addition, the training dataset should contain only few and easy maintainable predictors. To reinforce these demands, we extended the combination model presented in Schaumann et al. (2021) in order to improve its forecast quality and to make it more suitable for an operational setting. Furthermore, we evaluated the forecast quality of the hyper-parameter optimized combination model, when trained on a new high-resolution dataset. This dataset consists, on the one hand, of forecasts of DWD's ensemble-based precipitation nowcasting algorithm STEPS-DWD (Reinoso-Rondinel et al. 2022) and, on the other hand, of ensemble forecasts produced by an experimental setup of the operational high-resolution short-term NWP model ICON-D2.

The validation results for the new dataset show that the combination model and its modification achieve similar scores as for the previously considered dataset (Schaumann et al. 2021). More precisely, we were able to show that our C^3-models are indeed consistent over the whole range of threshold exceedances considered in this study. The forecasts represent an optimal combination of the input forecasts of STEPS-DWD and ICON-D2, which is indicated by a higher Brier skill score over all thresholds and lead times. The impact of spatial smoothing caused by convolutions is reduced by the C^3-models. That is effected, first, due to the utilization of probabilities based on hourly rainfall amount and, second, due to the forecast calibration. The reliability diagrams of the combination models are well-calibrated for all lead times and at least for the two lowest thresholds. The only diagrams affected by the smoothing mentioned above are those of the thresholds of 1 mm and 2 mm, for the C^3-model at +3 h and for the C^3_{LT1}-model at +6 h. However, in case of the C^3-model, this may be attributed to the low number of triangular functions.

In an operational setting robust and interpretable forecasts are important, i.e., a forecast model should not only achieve high aggregate validation scores, but also produce spatially and temporally consistent forecasts. For this, we investigated the performance of both initial models and the combination models by considering spatially resolved validation scores, to see how well each model performs at single grid points. The spatially resolved scores of bias and BSS reveal typical shortcomings of radar measurements and radar compositing, e.g., range attenuation that was not corrected due to the operation of a single-polarization radar, beam blocking, and temporary radar outages. However, the resulting spatial patterns are also visible in those results of the C^3-models, indicating that these deficits are not learned by the latter models, since these deficits are also present in the ground truth.

Finally, we considered the flip-flop index as a measure of temporal consistency. The obtained results show that both combination models produce forecasts with spatially more homogeneous validation scores and an improved flip-flop score. However, some spatial artifacts remain along the boundaries of radar coverage areas, which is likely due to the radar composite being used as ground truth. A possible alternative ground truth for verification could be station measurements. Moreover, we tested a modification (C^3_{LT1}) of the C^3-model which led to increased sharpness and a slight improvement of the flip-flop score over the C^3-model.

b. Outlook

The current combination model produces probabilities for the exceedance of thresholds at single grid points. However, for weather warnings it would be useful to predict probabilities for the exceedance of thresholds within predefined areas (e.g. river basins or municipal territories). As a next step we will investigate how the current combination model can be modified in order to predict such area-dependent exceedance probabilities.

Additionally, we will extend the underlying dataset by the winter months 2021/2022 to investigate the performance of the combination model for different seasons, and whether additional predictors like orography, wind information, local forecast variance or ensemble spread improves the combined forecast.

Acknowledgement

We would like to thank Susanne Theis for valuable comments and suggestions which helped us to design and perform this study. Furthermore, We are grateful to three anonymous reviewers for their helpful comments on the manuscript.

For the implementation and training of the C^3-model the Tensorflow library was used.

Data Availability Statement

The dataset utilized in the present study consists of ICON-D2 and STEPS-DWD forecasts both in an experimental non-operational setup. Therefore, the dataset is not publicly available until the forecast systems are operational. Afterwards, the data can be found at https://opendata.dwd.de. The code basis is an essential point of our current research topic and, therefore, we do not want to publish the code at this stage. However, at a further stage publishing the code is possible.

References

Armstrong, J. S., and M. C. Grohman, 1972: A comparative study of methods for long-range market forecasting. *Management Science*, **19 (2)**, 211–221.

Atencia, A., and I. Zawadzki, 2014: A comparison of two techniques for generating nowcasting ensembles. Part I: Lagrangian ensemble technique. *Monthly Weather Review*, **142 (11)**, 4036–4052.

Ben Bouallègue, Z., and D. S. Richardson, 2022: On the ROC Area of Ensemble Forecasts for Rare Events. *Weather and Forecasting*, **37 (5)**, 787–796.

Bick, T., and Coauthors, 2016: Assimilation of 3D radar reflectivities with an ensemble Kalman filter on the convective scale. *Quarterly Journal of the Royal Meteorological Society*, **142 (696)**, 1490–1504.

Bouttier, F., and H. Marchal, 2020: Probabilistic thunderstorm forecasting by blending multiple ensembles. *Tellus A*, **72 (1)**, 1–19.

Bowler, N. E., C. E. Pierce, and A. W. Seed, 2006: STEPS: A probabilistic precipitation forecasting scheme which merges an extrapolation nowcast with downscaled NWP. *Quarterly Journal of the Royal Meteorological Society*, **132 (620)**, 2127–2155.

Brunet, G., S. Jones, and P. M. Ruti, 2015: *Seamless Prediction of the Earth System: from Minutes to Months*. World Meteorological Organization, iSBN: 978-9263111562.

Brunet, G., and Coauthors, 2010: Collaboration of the weather and climate communities to advance subseasonal-to-seasonal prediction. *Bulletin of the American Meteorological Society*, **91 (10)**, 1397–1406.

Cuomo, J., and V. Chandrasekar, 2021: Use of deep learning for weather radar nowcasting. *Journal of Atmospheric and Oceanic Technology*, **38 (9)**, 1641–1656.

Ehret, U., 2010: Convergence index: A new performance measure for the temporal stability of operational rainfall forecasts. *Meteorologische Zeitschrift*, **19**, 441–451.

Foresti, L., M. Reyniers, A. Seed, and L. Delobbe, 2016: Development and verification of a real-time stochastic precipitation nowcasting system for urban hydrology in Belgium. *Hydrology and Earth System Sciences*, **20 (1)**, 505.

Foresti, L., and A. Seed, 2014: The effect of flow and orography on the spatial distribution of the very short-term predictability of rainfall from composite radar images. *Hydrology and Earth System Sciences*, **18 (11)**, 4671.

Germann, U., and I. Zawadzki, 2002: Scale-dependence of the predictability of precipitation from continental radar images. Part I: Description of the methodology. *Monthly Weather Review*, **130 (12)**, 2859–2873.

Golding, B., 1998: Nimrod: A system for generating automated very short range forecasts. *Meteorological Applications*, **5 (1)**, 1–16.

Griffiths, D., M. Foley, I. Ioannou, and T. Leeuwenburg, 2019: Flip-flop index: Quantifying revision stability for fixed-event forecasts. *Meteorological Applications*, **26 (1)**, 30–35.

Haiden, T., A. Kann, C. Wittmann, G. Pistotnik, B. Bica, and C. Gruber, 2011: The Integrated Nowcasting through Comprehensive Analysis (INCA) system and its validation over the Eastern Alpine region. *Weather and Forecasting*, **26 (2)**, 166–183.

Han, L., J. Sun, W. Zhang, Y. Xiu, H. Feng, and Y. Lin, 2017: A machine learning nowcasting method based on real-time reanalysis data. *Journal of Geophysical Research: Atmospheres*, **122 (7)**, 4038–4051.

Hazeleger, W., and Coauthors, 2012: EC-Earth V2. 2: description and validation of a new seamless earth system prediction model. *Climate dynamics*, **39 (11)**, 2611–2629.

Heizenreder, D., P. Joe, T. Hewson, L. Wilson, P. Davies, and E. de Coning, 2015: Development of applications towards a high-impact weather forecast system. *Seamless Prediction of the Earth*

System: From Minutes to Months (WMO-N 1156), G. Brunet, S. Jones, and P. M. Ruti, Eds., World Meteorological Organization, 419–443.

Hess, R., 2020: Statistical postprocessing of ensemble forecasts for severe weather at Deutscher Wetterdienst. *Nonlinear Processes in Geophysics*, **27**, 473–487.

Johnson, A., and X. Wang, 2012: Verification and calibration of neighborhood and object-based probabilistic precipitation forecasts from a multimodel convection-allowing ensemble. *Monthly Weather Review*, **140 (9)**, 3054–3077.

Kober, K., G. C. Craig, C. Keil, and A. Dörnbrack, 2012: Blending a probabilistic nowcasting method with a high-resolution numerical weather prediction ensemble for convective precipitation forecasts. *Quarterly Journal of the Royal Meteorological Society*, **138 (664)**, 755–768.

Nerini, D., L. Foresti, D. Leuenberger, S. Robert, and U. Germann, 2019: A reduced-space ensemble kalman filter approach for flow-dependent integration of radar extrapolation nowcasts and nwp precipitation ensembles. *Monthly Weather Review*, **147 (3)**, 987–1006.

Nicolis, C., R. A. Perdigao, and S. Vannitsem, 2009: Dynamics of prediction errors under the combined effect of initial condition and model errors. *Journal of the Atmospheric Sciences*, **66 (3)**, 766–778.

Palmer, T., F. Doblas-Reyes, A. Weisheimer, and M. Rodwell, 2008: Toward seamless prediction: Calibration of climate change projections using seasonal forecasts. *Bulletin of the American Meteorological Society*, **89 (4)**, 459–470.

Prudden, R., S. Adams, D. Kangin, N. Robinson, S. Ravuri, S. Mohamed, and A. Arribas, 2020: A review of radar-based nowcasting of precipitation and applicable machine learning techniques. Preprint, arXiv:2005.04988.

Pulkkinen, S., V. Chandrasekar, A. von Lerber, and A.-M. Harri, 2020: Nowcasting of convective rainfall using volumetric radar observations. *IEEE Transactions on Geoscience and Remote Sensing*, **58 (11)**, 7845–7859.

Reinoso-Rondinel, R., M. Rempel, M. Schultze, and S. Trömel, 2022: Nationwide radar-based precipitation nowcasting - A localization filtering approach and its application for Germany.

IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, **15**, 1670–1691.

Richardson, D. S., H. L. Cloke, and F. Pappenberger, 2020: Evaluation of the consistency of ECMWF ensemble forecasts. *Geophysical Research Letters*, **47 (11)**, e2020GL087 934.

Ruth, D. P., B. Glahn, V. Dagostaro, and K. Gilbert, 2009: The performance of mos in the digital age. *Weather and Forecasting*, **24 (2)**, 504–519.

Ruti, P. M., and Coauthors, 2020: Advancing research for seamless Earth system prediction. *Bulletin of the American Meteorological Society*, **101 (1)**, E23–E35.

Schaumann, P., M. de Langlard, R. Hess, P. James, and V. Schmidt, 2020: A calibrated combination of probabilistic precipitation forecasts to achieve a seamless transition from nowcasting to very short-range forecasting. *Weather and Forecasting*, **35 (3)**, 773–791.

Schaumann, P., R. Hess, M. Rempel, U. Blahak, and V. Schmidt, 2021: A calibrated and consistent combination of probabilistic forecasts for the exceedance of several precipitation thresholds using neural networks. *Weather and Forecasting*, **36**, 1076–1096.

Schraff, C., H. Reich, A. Rhodin, A. Schomburg, K. Stephan, A. Perianez, and R. Potthast, 2016: Kilometre-scale ensemble data assimilation for the COSMO model (KENDA). *Quarterly Journal of the Royal Meteorological Society*, **142 (696)**, 1453–1472.

Seed, A. W., 2003: A dynamic and spatial scaling approach to advection forecasting. *Journal of Applied Meteorology*, **42 (3)**, 381–388.

Seed, A. W., C. E. Pierce, and K. Norman, 2013: Formulation and evaluation of a scale decomposition-based stochastic precipitation nowcast scheme. *Water Resources Research*, **49 (10)**, 6624–6641.

Steinert, J., P. Tracksdorf, and D. Heizenreder, 2021: Hymec: Surface Precipitation Type Estimation at the German Weather Service. *Weather and Forecasting*, **36 (5)**, 1611–1627.

Stephan, K., S. Klink, and C. Schraff, 2008: Assimilation of radar-derived rain rates into the convective-scale model COSMO-DE at DWD. *Quarterly Journal of the Royal Meteorological Society*, **134 (634)**, 1315–1326.

Ukkonen, P., A. Manzato, and A. Mäkelä, 2017: Evaluation of thunderstorm predictors for Finland using reanalyses and neural networks. *Journal of Applied Meteorology and Climatology*, **56 (8)**, 2335–2352.

Vannitsem, S., and Coauthors, 2021: Statistical postprocessing for weather forecasts–review, challenges and avenues in a big data world. *Bulletin of the American Meteorological Society*, **102 (3)**, E681–E699.

Venugopal, V., E. Foufoula-Georgiou, and V. Sapozhnikov, 1999: Evidence of dynamic scaling in space-time rainfall. *Journal of Geophysical Research*, **104 (D24)**, 31 599–31 610.

Zängl, G., D. Reinert, P. Rípodas, and M. Baldauf, 2015: The ICON (ICOsahedral Non-hydrostatic) modelling framework of DWD and MPI-M: Description of the non-hydrostatic dynamical core. *Quarterly Journal of the Royal Meteorological Society*, **141 (687)**, 563–579.

Zawadzki, I., J. Morneau, and R. Laprise, 1994: Predictability of precipitation patterns: An operational approach. *Journal of Applied Meteorology*, **33 (12)**, 1562–1571.

Zsoter, E., R. Buizza, and D. Richardson, 2009: "Jumpiness" of the ECMWF and Met Office EPS control and ensemble-mean forecasts. *Monthly Weather Review*, **137 (11)**, 3823–3836.