A calibrated combination of probabilistic precipitation forecasts to achieve a seamless transition from nowcasting to very-short-range forecasting

Peter Schaumann*, [1] and Mathieu de Langlard[1] and Reinhold Hess[2] and Paul James[2] and Volker Schmidt[1].

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

[2]*Deutscher Wetterdienst, Offenbach, Germany*

Corresponding author address: Institute of Stochastics, Ulm University, Helmholtzstr. 18, 89069 Ulm, Germany.

E-mail: peter.schaumann@uni-ulm.de, mathieu.de-langlard@uni-ulm.de, reinhold.hess@dwd.de, paul.james@dwd.de and volker.schmidt@uni-ulm.de.

ABSTRACT

In this paper, a new model for the combination of two or more probabilistic forecasts is presented. The proposed combination model is based on a logit transformation of the underlying initial forecasts involving interaction terms. The combination aims at approximating the ideal calibration of the forecasts which is shown to be calibrated and to maximize the sharpness. The proposed combination model is applied to two precipitation forecasts, Ensemble-MOS and RadVOR, which were developed by Deutscher Wetterdienst. The proposed combination model shows significant improvements in various forecast scores for all considered lead times compared to both initial forecasts. In particular, the proposed combination model is calibrated, even if both initial forecasts are not calibrated. It is demonstrated that the method enables a seamless transition between both initial forecasts across several lead times to be created. Moreover, the method has been designed in such a way that it allows for fast updates in nearly real time.

1. Introduction

In many situations, it is possible to have access to several probabilistic forecasts of the same event (Clemen 1989; Graham 1996; Ariely et al. 2000; Winkler and Poses 1993). As these forecasts might be provided by independent models, non negligible differences can be observed. It is then necessary to find a combination of all forecasts for decision makers. Keeping the probabilistic forecast that performs best for some specific scores, thus dropping the others, is not an optimal choice. It is sometimes worth keeping the information of relatively poor probabilistic forecasts regarding these same specific scores, provided there is some degree of statistical independence between the forecasts.

Recently, the rise of Artificial Neural Networks (ANN) for making predictions in various fields has also emphasized the power of forecast combination techniques. It can be observed for various Kaggle challenges (Pavlyshenko 2018) that the most performant ANN architectures (*i.e.* having the highest generalization capability) are actually aggregations of several individual ones (Chollet 2017). In the field of weather forecasting, the performance of aggregation methods has long been investigated and highlighted (Sanders 1963; Bosart 1975; Vislocky and Fritsch 1995; Baars and Mass 2005; Hamill et al. 2008; Ranjan and Gneiting 2010; Gneiting et al. 2013). It is therefore legitimate to wonder whether there is an efficient strategy to aggregate probabilistic forecasts in order to capture most of the relevant features of the individual ones.

Several methods for combining probabilistic forecasts have been proposed in the literature. They either combine subjective forecasts made by meteorologists or objective ones from Numerical Weather Prediction (NWP) models. Most of these techniques rely on a linearly weighted average of the probabilistic forecasts. For example, Sanders (1963) has suggested to use the equally weighted average of twelve subjective probabilistic forecasts as a combination method. In this

case-study, it has been shown that this new aggregated probabilistic forecast had a positive Brier skill score relative to the climatological forecast, but, more surprisingly, relative to the best forecaster of the group as well. Vislocky and Fritsch (1995) investigated the average of two post-processed (with a Model Output Statistics (MOS) method) objective forecasts derived from two different high resolution models. They concluded that the combination product had a higher skill than the two individual MOS forecasts, allowing one to provide reliable forecasts for higher lead times regarding temperature, wind speed, probability of cloud and precipitation amount. Other works related to a linearly weighted average aggregation of probabilistic forecasts include Winkler et al. (1977), Gyakum (1986), Baars and Mass (2005), Hamill et al. (2008).

Ranjan and Gneiting (2010) have proved that a linearly weighted combination of distinct probabilistic forecasts is not the best combination strategy. In general it leads to uncalibrated forecasts, regardless of whether the underlying individual forecasts are calibrated or not. This important theoretical result does not state that such a combination would necessarily decrease the forecast skill of the combined forecasts below the forecast skill of the initial forecasts, but rather that it is sub-optimal and can potentially be improved by using a non-linear transformation instead. Thus, it does not contradict the other empirical results described in the previous paragraph. As a consequence, Ranjan and Gneiting (2010) suggested a beta-transformed linearly weighted combination of several forecasts. Their numerical results have highlighted some significant improvements in the reliability and sharpness of the forecasts compared to the classic linearly weighted average. The beta-transformed linearly weighted combination has later been adapted in Bassetti et al. (2018) for the combination of predictive probability distributions. For a comparison of methods for the combination of predictive distributions see Baran and Lerch (2018).

Following Ranjan and Gneiting's work, the goal of the present paper is twofold: 1) to give another theoretical interpretation of calibrated and sharp combined probabilistic forecasts, and 2) to

propose a non-linear combination that enables one to significantly increase the forecast quality for a dichotomous event. The dichotomous event considered in this paper is that of precipitation above 0.1mm per hour. The suggested model is applied to two forecasts (called Ensemble-MOS and RadVOR) developed by *Deutscher Wetterdienst* (DWD), Germany's National Meteorological Service. Ensemble-MOS is a short-term probabilistic forecast (up to 21 hours), while RadVOR provides predictions for up to two hours. Generally, RadVOR has better forecast scores for very short lead times, whereas for longer lead times Ensemble-MOS forecasts are preferably used. The proposed combination model is aimed at capturing most information of the two initial forecasts while achieving a seamless transition between both precipitation forecasts across several lead times, see Bowler et al. (2006), Golding (1998), Kober et al. (2012).

The rest of the paper is organized as follows. In Section 2, the Ensemble-MOS and RadVOR forecast data is described. A method is proposed for the transformation of the deterministic RadVOR forecasts into point probabilities, see Theis et al. (2005). Moreover, rain gauge adjusted radar precipitation measurements are presented as they are used for validation purposes. In Section 3, the notions of calibration and sharpness are defined. Some theoretical considerations on calibrated and sharp probabilistic forecasts are also presented. In Section 4, our model is described for the combination of two probabilistic forecasts. Then, in Section 5, the proposed model is numerically validated. Finally, in Section 6 it is shown that the developed method can also be applied to the combination of so-called area probabilities. The paper closes with a conclusion and an outlook to some future developments in Section 7.

2. Data

a. Ensemble-MOS

Ensemble-MOS of DWD is a model output statistics (MOS) system specialized for the optimization and calibration of probabilistic forecasts based on ensemble systems. In this paper it is applied to COSMO-DE-EPS, the ensemble system of the high-resolution convection-permitting model COSMO-DE of DWD. Ensemble products as mean and standard deviation for a set of model fields are used as predictors in multiple linear and logistic regressions against conventional synoptic observations including rain gauges, especially for precipitation forecasts. Ensemble-MOS forecasts based on 5 years of training data (2011-2015) were used in order to provide precipitation forecasts from May to July 2016 with lead times from 1h to 21h on a 20km $\times$ 20km grid.

b. RadVOR

1) DETERMINISTIC FORECASTS

DWD runs an operational quantitative precipitation estimation (QPE) system, called RADOLAN (Weigl and Winterrath 2010). The DWD radar network provides the basis for optimized national composites of current radar reflectivities to be generated on a 5-minute update cycle. RADOLAN then combines empirical Z-R relationships with real-time rainfall gauge measurements from the synoptic station network to yield a calibrated best estimate of current rainfall rates.

For the purposes of providing forecasts and warnings of potential heavy rainfall on nowcasting timescales, DWD has developed a follow-on operational system, called RadVOR (Winterrath et al. 2012), which gives quantitative rainfall forecasts (QPF) for the next two hours with an update cycle of 5 minutes. The rainfall estimates from RADOLAN are extrapolated forwards in time with the

aid of an optimized rainfall displacement vector field. This field is calculated via a mapping of precipitation patterns in successive image data, taking different spatial motion scales into account and using satellite motion vectors to add stability, for example in areas where no precipitation is present. RadVOR provides moving rainfall estimates in 5-minute forecast steps on a 1x1 km grid over the whole territory of Germany as well as summing up rainfall totals for the first and second forecast hours.

2) TRANSFORMATION OF DETERMINISTIC FORECASTS TO PROBABILISTIC FORECASTS

A method is outlined to convert the deterministic RadVOR forecasts to hourly point probabilities on the same grid as the Ensemble-MOS forecasts in order to unify the format of both forecasts.

Aggregation of RadVOR forecasts in time:

While Ensemble-MOS provides predictions for time intervals of 60 minutes, RadVOR has a forecast interval of 5 minutes length. In order to unify the forecast lengths, all RadVOR forecasts within one hour are aggregated by summation. The result is a deterministic prediction of precipitation amounts for one complete hour.

Recall that in this paper lead times up to $+6$ hours are considered, although RadVOR only produces forecasts up to $+2$ hours. Thus, when determining RadVOR forecasts for lead times above $+2$ hours, the last available 5-minute prediction is inserted repeatedly. This means that for periods with a lead time between $+2$ and $+3$ hours, some of the 5-minute predictions are identical. Aggregated predictions for periods with a lead time larger than $+3$ hours are all identical and consist of the sum of 12 identical 5-minute predictions. It is to be expected that this approach (compared to an aggregation of 12 different 5-minute intervals) leads to concentrated peaks of precipitation and therefore leads to a biased forecast.

It has been tested how well the hourly forecasts would perform if the last 12 available 5-minute forecasts would be used repeatedly instead for higher lead times. This alternative approach leads to a smaller bias of -0.005 for lead times from $+2h$ to $+6h$, but the Brier skill score and the reliability are significantly worse.

It should be noted that the development of a more sophisticated transformation from deterministic to probabilistic forecasts is outside the scope of this paper. The transformed RadVOR forecast merely serves as uncalbirated initial forecast for the proposed combination method. Furthermore, the decision to consider lead times longer than $+2h$ was made once it turned out that the combination of both forecasts is feasible for up to $+6h$. The RadVOR forecast still holds some valuable information for higher lead times, even if a persistence based extrapolation for up to $+6h$ seems not completely satisfactory from a meteorological perspective.

Local averaging:

In order to transform the hourly aggregated RadVOR forecasts into probabilistic weather forecasts, a similar approach as in Theis et al. (2005) is used. Recall that Ensemble-MOS predicts the likelihood that precipitation at a certain point within an hour exceeds a given threshold. In the present paper forecasts for the threshold of 0.1 mm are considered. To transform the aggregated RadVOR forecasts accordingly, the predicted hourly precipitation amounts are binarized for the threshold 0.1 mm. This means that precipitation amounts equal or larger than 0.1 mm are set equal to 1, while precipitation amounts below this threshold are set equal to 0. Let $V(r')$ denote this binarized value for a grid point $r' \in R'$ on the 1km$\times$1km grid R' and let R denote the 20km$\times$20km grid. Finally, a weighted average $\overline{V}(r)$ of the binarized values is calculated for each $r \in R$ using the formula

$$\overline{V}(r) = \frac{1}{\sum_{r' \in R'} w(r,r')} \sum_{r' \in R'} w(r,r') V(r') \tag{1}$$

with weights $w(r,r') = \|r-r'\|^{-1.75}$, where $\|\cdot\|$ is the Euclidean distance. The exponent -1.75 has been chosen empirically from the set $\{-1,-1.25,\ldots,-2.75,-3\}$, because it achieved the best reliability for the lead time $+1$h. The resulting average is considered as the probability for the exceedance of 0.1mm of precipitation. Since the influence of $V(r')$ on $\overline{V}(r)$ becomes negligible for larger distances between r and r', only grid points with $\|r-r'\| \leq 50$km are considered.

c. Calibrated hourly radar-measurements

In order to validate the results obtained in in this paper, rain gauge adjusted radar precipitation measurements are used. The measurements were made by the German operational radar network of DWD (Winterrath et al. 2012), which covers Germany with 16 radar sites that provide scans in intervals of 5 minutes.

The rate of precipitation is derived by transforming the measured radar reflectivities based on empirical reflectivity-precipitation rate (Z-R) relationships, whereas 0.1 mm/h of precipitation is the minimum amount which can be detected. To improve accuracy, the precipitation amounts are adjusted according to the measurements of about 1,300 rain gauges which are located at meteorological measurement sites. Finally, pixel artifacts, which may occur in radar scans, are removed by a clutter filter as proposed by Winterrath and Rosenow (2007).

3. Mathematical background

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be some abstract probability space, *i.e.*, Ω is a non-empty set describing all possible states of a certain system, $\mathscr{F}$ a σ-algebra of subsets of Ω and $\mathbb{P}$ a probability measure on $\mathscr{F}$. For instance, Ω can be the set of all possible meteorological scenarios for a given region.

a. Self-calibration as an optimal combination approach

Let P be a continuous random variable taking values in the unit interval $[0,1]$, and Y be a dichotomous random variable taking as values 1 with probability q and 0 with probability $1-q$, where $0 \leq q \leq 1$. The random variable P represents a probabilistic forecast for the event $Y = 1$, *i.e.*, that the amount of precipitation exceeds the threshold $T = 0.1$mm.

In this paper, the probabilistic forecast P is said to be calibrated if

$$\mathbb{P}(Y = 1 \mid P) = \mathbb{E}(Y \mid P) = P. \tag{2}$$

where $\mathbb{P}(Y = 1 \mid P)$ denotes the conditional probability that the event $Y = 1$ occurs, given the probabilistic forecast P. Analogously, $\mathbb{E}(Y \mid P)$ denotes the conditional expectation of Y given P. This notion of calibration means that the information delivered by the probabilistic forecast P is *reliable*, see also Murphy and Winkler (1977, 1987). A direct consequence of Eq. (2) is that on average the forecast provides the probability of appearance of the event $Y = 1$, *i.e.*, $\mathbb{E}(Y) = \mathbb{P}(Y = 1) = \mathbb{E}(P)$.

If P is uncalibrated, then

$$f(P) = \mathbb{E}(Y \mid P) \neq P, \tag{3}$$

where f is an unknown deterministic function. Besides, from basic properties of conditional expectation, the random variable $f(P)$ is itself calibrated (see the Appendix for some mathematical background). Naturally, $f(P)$ is called the *self-calibrated version* of P. More generally, if $P_1, ..., P_n$ are n probabilistic forecasts, then

$$f(P_1, ..., P_n) = \mathbb{E}(Y \mid P_1, ..., P_n) \tag{4}$$

is the self-calibrated version of the aggregation of the n probabilistic forecasts.

The notion of calibration is an important property that a probabilistic forecast should exhibit. However, the notion of calibration is not sufficient for characterizing the skill of a forecast. For

example, the climatological forecast P, which predicts the average probability of precipitation only, is perfectly calibrated but not a useful prediction. Therefore, assuming calibration, the notion of *sharpness* makes it possible to discriminate the useful informative forecasts (Gneiting et al. 2007).

The sharpness is defined as the variance $\mathrm{Var}(P)$ of the forecast P and corresponds to the dispersion of the forecast from the forecast average. The sharper a forecast, the more P takes values close to 0 and 1; hence, the higher the variance. Note that sharpness alone is not a measure for forecast quality, since sharpness is only a property of the distribution of the predicted probabilities but is not affected by how accurate these probabilities are.

The self-calibrated version $f(P)$ of P is the most sharp probabilistic forecast among all calibrated ones which depends on P in the sense that it is the solution of

$$f(P) = \underset{\substack{g \in G \\ \text{s.t: } \mathbb{E}(Y|g(P))=g(P)}}{\arg\max} \ \mathrm{Var}(g(P)), \tag{5}$$

where G is the set of deterministic functions $g : [0,1] \to [0,1]$ such that $g(P)$ is a well-defined random variable. The proof of Eq. (5) is given in the Appendix. This result generalizes naturally for the self-calibrated version $f(P_1,...,P_n)$ of several probabilistic forecasts $P_1,...,P_n$. Note that in Ranjan and Gneiting (2010) it has been proven that a linear combination of n forecasts given by $g(P_1,\ldots,P_n) = w_1P_1 + \ldots + w_nP_n$, where $w_1,\ldots,w_n$ are some weights, lacks calibration and sharpness compared to the self-calibrated version of the forecasts. Our approach is more general in that it combines the initial forecasts in a non-linear way and considers interactions between them.

Another fundamental property of the self-calibrated version of probabilistic forecasts is that it is the best approximation of Y with respect to the L_2-norm, *i.e.*,

$$f(P_1,...,P_n) = \underset{Z \text{ is } \sigma(P_1,...,P_n)\text{-measurable}}{\arg\min} \mathbb{E}\left((Z-Y)^2\right). \quad (6)$$

This property is due to the fact that the conditional expectation is the orthogonal projection of Y on the space of $\sigma(P_1,...,P_n)$-measurable random variables, where $\sigma(P_1,...,P_n)$ is the sub-σ-algebra of $\mathscr{F}$ generated by the random variables $P_1,...,P_n$. Eq. (6) means that f minimizes the expected Brier score (see Section 5) and also any strictly proper scoring rule as proven by Ranjan and Gneiting (2010).

For all of these reasons, the self-calibrated version of any set of probabilistic forecasts is the best combination method to employ. However, in general the self-calibrated version f of forecasts is unknown and therefore intractable : in practice it is not possible to have a closed-form formula for the function f (only the existence is ensured). Therefore, some parametric assumptions are usually made on f.

b. Parametric types of combination

The most commonly used approximation of f is the *linear pool* f_{LP} defined by

$$f_{\text{LP}}(P_1,...,P_n) = w_1P_1 + ... + w_nP_n, \quad (7)$$

where the weights $w_1,\ldots,w_n$ are such that $0 \leq w_i \leq 1$ and $w_1 + ... + w_n = 1$. This type of combination has been widely investigated in the literature, see Baars and Mass (2005), Bosart (1975), Genest and McConway (1990), Clemen and Winkler (1999). However, it has been shown by Ranjan and Gneiting (2010) that the linear pool is not optimal, even if the underlying forecasts are assumed to be calibrated (see Theorem 1 in their paper).

This is why Ranjan and Gneiting (2010) proposed a more complex parametric approximation as a combination model. They used a non-linear transformation of the linear pool, denoted by f_{BLP},

where

$$f_{\mathrm{BLP}}(P_1,...,P_n) = H_{\alpha,\beta}\left(f_{\mathrm{LP}}(P_1,...,P_n)\right). \tag{8}$$

The function $H_{\alpha,\beta}$ in Eq. (8) is the cumulative distribution function of the beta distribution with shape parameters $\alpha > 0$ and $\beta > 0$ defined by

$$H_{\alpha,\beta}(x) = \int_0^x t^{a-1}(1-t)^{b-1}dt, \text{ for all } x \in [0,1]. \tag{9}$$

It has been shown empirically in Ranjan and Gneiting (2010) that this non-linear transformation increases the reliability and the sharpness of the combined forecast compared to the linear pool and all initial forecasts $P_1,\ldots,P_n$.

In the present study, a new type of approximation is proposed for the self-calibrated version of two probabilistic forecasts that leads to a reliable and sharp forecast as highlighted in Section 5. The approximation is based on the logistic transformation of a non-linear combination of the underlying initial probabilistic forecasts with some interaction terms. This approximation of f is described in detail in the next section.

4. Generalized logit combination

The approximation of a conditional expectation of a dichotomous random variable Y given a set of predictors $P_1,...,P_n$ is often achieved with a so-called logit model (or logistic regression). In the literature, this model has been used for MOS methods in order to post-process ensemble members returned by a probabilistic forecast (Hamill et al. 2008; Wilks 2009; Ben Bouallègue 2013). In the present paper, a more general version of the logit model is proposed to approximate the self-calibrated version of a set of probabilistic forecasts. More specifically, the approximation is explicitly detailed for the combination of two probabilistic forecasts which generally give different predictions.

a. Logit combination with triangular functions

Given a set of predictors $P_1, ..., P_n$, the *standard logit model* is given as follows:

$$f_L(P_1, \ldots, P_n) = \sigma\left(a + \sum_{i=1}^{n} b_i P_i\right), \tag{10}$$

where $\sigma(x) = \frac{1}{1+\exp(-x)}$ is the sigmoid function and the coefficients a and $b_1, \ldots, b_n$ are some model parameters, Note that a is usually called the intercept of the model.

The initial forecasts P_i are not necessarily well calibrated. In such a situation, the standard combination model given by Eq. (10) may lead to an uncalibrated forecast as the sigmoid function of the simple linear pool is not flexible enough to compensate for the possible underestimation and overestimation of the P_i's (see Fig. 1 for an example of deviations). To mitigate these effects, each probabilistic forecast P_i is split into several predictors $\phi_0(P_i), \ldots, \phi_m(P_i)$, where the functions $\phi_0, \phi_1, \ldots, \phi_m$ are given by

$$\phi_j(x) = \max\left\{0, 1 - m|x - \frac{j}{m}|\right\}, \; x \in [0,1] \tag{11}$$

for all $j \in \{0, 1, \ldots, m\}$. These functions are called *triangular functions*. In Fig. 2 a set of triangular functions is shown for $m = 5$. Noticing that $\phi_0(x) + \ldots + \phi_m(x) = 1$ for all $x \in [0,1]$, the intercept coefficient becomes unnecessary and the logit model of Eq. (10) transforms into a more flexible model $f_{LT}(P_1, \ldots, P_n)$ based on the triangular functions $\phi_0, \ldots, \phi_m$:

$$f_{LT}(P_1, \ldots, P_n) = \sigma\left(\sum_{i=1}^{n} \sum_{0=1}^{m} b_{ij} \phi_j(P_i)\right). \tag{12}$$

For example, for $n = 1$, the logit combination model stated in Eq. (12) takes the following form:

$$f_{LT}(P_1) = \sigma(w_0 \phi_0(P_1) + \ldots + w_m \phi_m(P_1)), \tag{13}$$

where $w_1, \ldots, w_m$ are some parameters and the family of triangular functions $\phi_0, \phi_1, \ldots, \phi_m$ is constructed such a way that the expression $w_0 \phi_0(P_1) + \ldots + w_m \phi_m(P_1)$ can be considered to be a

piecewise linear interpolation between the points $(\frac{0}{m}, w_0), (\frac{1}{m}, w_1), \ldots, (\frac{m}{m}, w_m)$, which transforms the values of P_1 accordingly. In this way, the model given in Eq. (13) is able to compensate over- and underestimations for different values of P_1 at the same time.

b. Interaction terms

Consider the case of two initial probabilistic forecasts P_1 and P_2. Let m be the chosen number of triangular functions. Fig. 3 shows the effects of single triangular functions on the output of the combination model. The output of the combination model f_{LT} for the crossing points $(0.1, 0.1)$, $(0.1, 0.8)$,$(0.5, 0.1)$ and $(0.5, 0.8)$ in the bottom left subplot is fully determined by the coefficients of the four triangular functions. While there are four points and four coefficients, it is generally impossible to find a set of coefficients such that the model output for these four points matches with an arbitrary set of four probabilities, i.e., the model can choose the four coefficients so that the probabilities of only 3 of the 4 points are correctly predicted. See the Appendix for a proof. In order to be able to make correct predictions for all four points, the model needs more degrees of freedom. For this, some *interactions terms* of the forecasts P_1 and P_2 are considered, which consist of the four functions $\gamma_1, \gamma_2, \gamma_3, \gamma_4$ defined on $[0,1]^2$ by

$$\begin{aligned}
\gamma_1(p_1, p_2) &= \sqrt{p_1 p_2}, \\
\gamma_2(p_1, p_2) &= \sqrt{(1-p_1) p_2}, \\
\gamma_3(p_1, p_2) &= \sqrt{p_1 (1-p_2)}, \\
\gamma_4(p_1, p_2) &= \sqrt{(1-p_1)(1-p_2)}
\end{aligned}$$

for $p_1, p_2 \in [0,1]$.

Keeping the triangular functions considered in Eq. (12) and incorporating the interactions terms leads to the following *generalized logit combination model*:

$$f_{LTI}(P_1,P_2) = \sigma\left(\sum_{i=1}^{2}\sum_{j=0}^{m} a_{ij}\phi_j(P_i) + \sum_{i=1}^{4}\sum_{j=0}^{m} b_{ij}\phi_j(\gamma_i(P_1,P_2))\right), \tag{14}$$

where a_{ij} and b_{ij} are some model parameters. Thus, there are $6(m+1)$ parameters to be fitted.

In the upper right subplot of Fig. 3 three triangular functions for γ_1 are depicted. The triangular functions of the interaction terms allow the model to chose coefficients for the case when the two forecasts P_1 and P_2 predict both high probabilities (for γ_1), low probabilities (for γ_4), or make diverging predictions (for γ_2 and γ_3), namely the four corners of $[0,1]^2$.

It has to be emphasized that the model given in Eq. (14) creates a fine-tuned combination between P_1 and P_2 with interaction terms, but also enables to be corrected systematic unreliable forecasts as a MOS method would do. A numerical validation of the combination model proposed in Eq. (14) is performed in the next section.

5. Numerical validation

In this section, the performance of the combination model proposed in Eq. (14) is analyzed using several validation scores. In particular, the model given in Eq. (14) is compared to the initial probabilistic forecasts (RadVOR and Ensemble-MOS) and also to the standard logit combination model f_L given in Eq. (10).

a. Validation scores

Various forecast scores can be used in order to assess the accuracy and the skill of a forecast (Wilks 2006). The following validation scores are considered in this paper: bias, Brier score, Brier skill score, reliability, and reliability diagram.

1) BIAS

The *bias* of a probabilistic forecast P is defined as the expected difference between the forecast P and the dichotomous random variable Y with $\mathbb{E}(Y) = q$, *i.e.*,

$$\text{Bias}(P) = \mathbb{E}(P - Y) = \mathbb{E}(P) - q. \tag{15}$$

An accurate precipitation forecast P makes predictions with a bias close to 0, which indicates that the occurrence of rain is neither overestimated nor underestimated on average. As already mentioned in Section 3, a calibrated forecast P is necessarily unbiased.

2) BRIER SCORE AND BRIER SKILL SCORE

The *Brier score* is given by the expected squared error between the forecast P and the dichotomous random variable Y, *i.e.*,

$$\text{BS}(P) = \mathbb{E}((P - Y)^2). \tag{16}$$

It is a measure of accuracy that is sensitive to strong deviations of given forecasts to their actually observed counterparts.

Furthermore, in order to assess the skill of a forecast, the *Brier skill score* is often used. It is based on a comparison of the Brier score of the forecast and the one of a reference forecast P_{ref} used as a *benchmark*, *i.e.*,

$$\text{BSS}(P) = 1 - \frac{\text{BS}(P)}{\text{BS}(P_{ref})}.$$

In this paper, the average $P_{ref} = q$ for the selected period May to July 2016 of the occurrence of precipitation exceeding the threshold 0.1mm is considered as a reference forecast. Note that if the Brier score of the forecast is lower than that of the reference forecast, then the Brier skill score is positive. In this case, the proposed forecast is considered to be skillful.

3) RELIABILITY AND RELIABILITY DIAGRAM

The *reliability* score is considered as a measure of conditional bias. Assume that for the probabilistic forecast n predictions $p_1, ..., p_n$ are available, which correspond to n observations $y_1, ..., y_n$ of the considered event. Denote by $B_1, ..., B_I$ a partition of the unit interval $[0, 1]$ into I subintervals. Each partition component B_i contains N_i values of forecasts p_k. These forecast values correspond to the observations of the event y_k. By $\bar{p}_i$ the average of the forecasts within B_i is denoted and by $\bar{y}_i$ the relative frequency of the events which correspond to the forecasts within B_i, *i.e.*,

$$\bar{p}_i = \frac{1}{N_i} \sum_{k \in B_i} p_k, \tag{17}$$

$$\bar{y}_i = \frac{1}{N_i} \sum_{k \in B_i} y_k. \tag{18}$$

Then, the reliability is defined as

$$\mathrm{Rel}_I(P) = \frac{1}{n} \sum_{i=1}^{I} N_i (\bar{p}_i - \bar{y}_i)^2. \tag{19}$$

The *reliability diagram* is the graphical representation of the (p_k, y_k)-pairs. The deviation of the reliability diagram from the first bisector of the axes is a qualitative visualization of the reliability. For a quantitative assessment, each reliability diagram is enclosed in a band. The upper and lower end of the band are the 95% and 5% quantiles of the reliability diagrams for single locations.

b. Training and testing procedure

For the validation results presented in this section, each forecast has been trained and tested using a *rolling-origin with reoptimization scheme* initially proposed by Armstrong and Grohman (1972). During this procedure, the model is updated with new training data for each hourly step of the time series in chronological order. The point in time T, until which the model has been trained, is called the *forecasting origin* and represents the current time in an operational scenario. The forecasting origin splits the data into available data from the past (training set) and unavailable data from the future (the test set). For each training step, the forecasting origin is moved one hour forwards in time and the model is updated with the new data that became available for training. The update means that the optimization procedure is run with the new available data. At the forecasting origin T, the model makes predictions for the future time interval $[T+L-1, T+L]$, where L is the chosen lead time in hours. The forecasting origin T is rolled over until $T+L \leq M$, where M is the final time of the data set. As the forecast quality of the initial forecasts (here RadVOR and Ensemble-MOS) are likely to depend on the lead times, each model has been trained independently for the considered lead times. Therefore, it is possible to assess the accuracy and the skill of the combination model with respect to the lead times.

The rolling-origin with reoptimization approach enables us to have more testing data when the data set is not too large and quantify the amount of data required for the training (Tashman 2000). The next section provides the results of an experimental study of the training procedure for the proposed combination model f_{LTI} in Eq. (14).

c. Evaluation of the fitted model

Before fitting the model to a given data set, two important parameters, called *hyperparameters*, need to be fixed:

1. the learning rate η used in the optimization algorithm for updating the model parameters, where the so-called *stochastic gradient descent algorithm* is considered in the present paper, see also Bottou (2010). The learning rate determines the magnitude of change of the parameters in each training step: a too high learning rate value may cause the algorithm to miss the global minimum (or a desirable local minimum), but a too small value may result in the algorithm taking long to converge or even getting stuck in an undesirable local minimum (see also Goodfellow et al. (2016) for further details),

2. the number m of triangular functions $\phi_1, \ldots, \phi_m$ for the proposed combination model.

In Fig. 4 the effect of η and m on the validation scores is shown. It seems that models with a higher number of triangular functions also require a higher learning rate. However, there does not seem to be a combination of hyperparameters that is superior to all others, especially if the same set of hyperparameters is chosen for all lead times. For the results presented in this paper, the hyperparameters of the model f_{LTI} have been set to $\eta = 0.0005$ and $m = 10$, which perform well for all considered forecast scores and all considered lead times. While there are other hyperparameter configurations with a similar performance, it has to be taken into account that the number of model weights increases with an increase of m and therefore should be chosen as low as possible.

For the standard logit combination model f_L the appropriate learning rate η has been determined in a similar way, by comparing the Brier skill scores for different learning rates, where $\eta = 0.0025$ performed best for short lead times, $\eta = 0.001$ for the mid range lead times and $\eta = 0.0005$ for long lead times. Since the differences were not significant (below 0.001), $\eta = 0.001$ was chosen for all lead times.

Once the hyperparameters were fixed, the models were fitted to the data using the rolling-origin with reoptimization procedure (see Section 5b). Fig. 5 visualizes the output of the fitted model

f_{LTI} and the corresponding observed probabilities. Notice that the proposed combination model gives more significance to forecasts provided by RadVOR for short lead times, while Ensemble-MOS is given more emphasis for longer lead times. This is in accordance with the validation scores since the RadVOR forecasts perform better than Ensemble-MOS forecasts at shorter lead times and worse for the longer lead times (see Fig. 6 and 1).

Fig. 7 depicts the distribution of the parameters a_{ij} and b_{ij} of the fitted combination model f_{LTI} introduced in Eq. (14) for the months of June (in red) and July (in blue) with violin plots. In this model, the initial probabilistic forecasts P_1 and P_2 (based on Ensemble-MOS and RadVOR) are split into 11 triangular functions $\phi_0, \ldots, \phi_{10}$, resulting in 11 parameters for each probabilistic forecast. Also, each interaction term $\gamma_1, \gamma_2, \gamma_3$ and γ_4 is decomposed into 11 triangular functions. For each value $x \in \{0, 0.1, \ldots, 0.9, 1\}$ on the x-axis, there is a triangular function ϕ, with $\phi(x) = 1$, the corresponding parameter of which is depicted at x in Fig. 7. For example for the value $x = 0$ regarding the RadVOR column, the violin plots in blue, respectively in red, can be seen as the influence of RadVOR predictions close to the value $x = 0$ on the combination model for the month of June, respectively of July. For the lead time $+1$h the RadVOR parameters range from -2 to $+1.5$, while the Ensemble-MOS parameters are between -0.5 and 0.5. Therefore, the predictions based on RadVOR have a larger influence on the combined forecast. With increasing lead times, Ensemble-MOS parameters spread out further and RadVOR parameters move closer to 0. These observations are consistent with those made regarding Fig. 5. Moreover, the parameters for Ensemble-MOS and γ_1 at $x = 1$ are close to zero because Ensemble-MOS made almost no predictions close to 1 (see the bar plots in Fig. 1 and data plots in Fig 5).Therefore, these parameters get seldomly updated and stay close to 0. It is notable that most parameters show a similar distribution for both months of June and July. Data for the month of May has been omitted due to the warm-up period at the beginning of the training, which leads to different parameter distributions

for May in comparison to June and July. Also, it can be seen that the variance of the parameter distribution increases for longer lead times. This is probably due to increased forecast errors in the initial forecasts. Note that if all 11 weights of a predictor are arranged on a line, then the triangular functions mimic the behavior of a standard logit combination model with one parameter for each initial predictor. However, the ability to choose parameters in a non-linear way leads to a more general and flexible combination model.

The interaction terms γ_1 and γ_4 take values close to 1 if both initial forecasts agree. In Fig. 7 it can be seen that if both initial forecasts predict precipitation, γ_1 further increases the predicted probability of the model, while if both initial forecasts predict no precipitation, γ_4 decreases the predicted probability further. γ_2 takes values close to 1 if Ensemble-MOS predicts no precipitation, but RadVOR does. For lower lead times, when RadVOR has a high forecast skill, γ_2 further increases the predicted probability of the model. For higher lead times and a lower forecast skill of RadVOR, the weights of γ_2 move closer to zero. Similarly the slope of γ_3 changes with increasing lead time according to which of the initial forecasts has a higher forecast skill.

The bias, Brier skill score, reliability and sharpness of the initial forecast, of the standard logit combination model f_L and of the proposed combination model f_{LTI} are shown in Fig. 6. The box plot diagrams represent the variability of the daily scores depending on lead time. They measure the consistency of the probabilistic forecasts from day-to-day predictions: the wider a box plot diagram is, the less consistent is the model. The continuous lines represent the validation scores over all locations and points in time of the data set. Note that the Brier skill score of 3 months is not equal to the average daily Brier skill score, which is more sensitive to days with a low Brier skill score. The overall scores for the combination model f_{LTI} are significantly better than those for the initial probabilistic forecasts with respect to the Brier skill score and the reliability. Ensemble-MOS shows little increasing bias, RadVOR a negative Bias of -2% and the combination models

are almost perfect for the 3 month average. Moreover, the daily predictions of the proposed model are more consistent than the initial forecasts. Besides, the proposed combination model preserves the sharpness for short lead times, but decreases it for longer lead times. Notice that all the scores of f_{LTI} are also improved compared to the standard logit combination model. In order to see the effect of interaction terms on the validation scores, the forecasts have been combined with a model of type f_{LT}, which extends the logistic regression model f_L with triangular functions only. The results show that f_{LTI} compared to f_{LT} (not shown here) has improved bias, reliability and sharpness.

Reliability diagrams are shown for these probabilistic forecasts in Fig. 1. The histograms represent the empirical distributions of the probabilistic forecasts. It seems that the combination model f_{LTI} is significantly more reliable for all lead times compared to the initial probabilistic forecasts and to the standard logit combination model. Fig. 6 and 1 highlight that the f_{LTI} combination model has a higher accuracy and skill than the initial probabilistic forecasts without impacting too much of the sharpness.

For the results presented in this paper, the combination model f_{LTI} has been trained on all point probabilities regardless of their corresponding location. Therefore the combination model can not correct local errors, which affect only a subset of locations. To assess how well the combination model performs for single stations, the considered forecast scores for each location are shown in Figs. 8, 9 and 10. Especially for the bias and the Brier skill score local differences can be observed for the combination model. However these differences seem to occur already in the initial forecasts and are not introduced by the combination model. In Fig. 10 the local reliability of the combination model is much more homogeneous than for both initial forecasts.

In Fig. 11 the initial and combined point probabilities are illustrated for one hour to showcase the seamless transition between both initial forecasts.

d. Runtime of the fitted model

In addition to validation scores, the runtime of a model is critical for operational use, especially if the initial forecasts have a fast update cycle of a few minutes like RadVOR. To benchmark the runtime of the proposed combination model f_{LTI}, the model was run on an Intel Core i7-860 (2.8 Ghz).

In order to combine 2210 hourly forecasts for approximately 1370 locations and 8 lead times, it took 41 minutes and 11 seconds to combine both considered forecasts, which corresponds to 1.118 seconds per hourly forecast. This includes reading the initial forecasts from a file, making a prediction for each location, saving the new predictions to a file and updating the model parameters with the new observations. The transformation of the RadVOR forecasts has not been considered in this evaluation, since the transformation is independent of the combination itself and does not affect the runtime in the general use case of the proposed model f_{LTI}.

Note that the model only requires the most recent information of the last hour to make the next prediction and to update the model parameters, which results in the short runtime and also in a low memory use.

6. Application to area probabilities for warning events

In this section the wide applicability of the approach proposed in this paper for the calibrated combination of probabilistic precipitation forecasts is demonstrated. More precisely, we show that our approach can also be used for the calibrated combination of so-called area probabilities. Note that most NWP models generate predictions for single points on a certain grid. This is also the case for RadVOR and Ensemble-MOS. In Kriesche et al. (2015), a stochastic geometry model has been introduced, which calculates area probabilities based on point probabilities. This model was developed for the generation of weather warnings. For instance, in order to predict the likelihood

of flooding, the probability of precipitation within the catchment area of a river is of interest, without knowing the exact location of the precipitation event. Similarly, emergency forces might have an interest in the area probability for critical weather events in their area of responsibility.

In our case, area probabilities can be defined as the probability of precipitation exceeding the threshold 0.1mm in at least one point within a certain fixed area A. From this definition, it follows that area probabilities of a given weather event are at least as large as the probabilities for single points or arbitrary subsets within A. Formally, the area probability $p(A)$ for the occurrence of precipitation anywhere inside A has the following representation, see e.g. Hess et al. (2018):

$$p(A) = 1 - \exp\left(-\sum_{s \in S} a(s) \nu_2 \left((A \oplus b(o,r)) \cap V(s) \right) \right), \tag{20}$$

where S is the set of points for which point probabilities are given, $V(s)$ is the Voronoi cell corresponding to location s, $a(s)$ is a model parameter representing the number of precipitation cells per unit area in $V(s)$. Furthermore, $\nu_2(G \oplus b(o,r))$ is the area of the dilated set $A \oplus b(o,r)$ where $A \oplus b(o,r)$ denotes the Minkowski sum of A and the disk $b(o,r)$ which is centered at the origin and has some radius $r > 0$ (Chiu et al. 2013). Note that the model parameters r and $a(s)$ for all $s \in S$ are estimated on the basis of corresponding point probabilities. For further details, we refer to Kriesche et al. (2015, 2017) .

In principle, combined area probabilities can be computed in two different ways. Namely, they can be computed

1. based on already combined point probabilities (Method 1) ;
2. for point probabilities of each initial forecast and then combined by the proposed combination model f_{LTI} (Method 2).

In Fig. 12 the validation scores for area probabilities based on RadVOR, Ensemble-MOS and their combination are compared, where the area probabilities for Ensemble-MOS and RadVOR

show similar behavior as the corresponding point probabilities in Fig. 6. Based on these forecast scores, Fig. 12 shows that Method 2 leads to a much smaller bias and better reliability than Method 1, whereas the BSS does not show any significant difference. Thus, when computing calibrated area probabilities, Method 2 described above should be used.

7. Conclusion

The combination model presented in this paper for combining probabilistic forecasts demonstrates significant improvements in forecast accuracy, skill and consistency with respect to all considered forecast scores. The forecast scores show even a large improvement for lead times where currently no RadVOR forecasts are available. Both the conversion of deterministic RadVOR predictions to probabilistic forecasts and the fitting of the proposed combination model are computationally rather cheap and, therefore, they allow for a seamless update of Ensemble-MOS forecasts.

Furthermore, the method has been applied to the combination of area probabilities, which can be used for warning events. The computation of area probabilities is based on a stochastic geometry model using point probabilities. The proposed method has been used to highlight that area probabilities should be computed from the point probabilities first and then combined with the combination model.

The combination model has not been applied to thresholds other than 0.1 mm yet. It is likely that a model trained for some threshold would not yield satisfactory results if it were applied to forecasts of another threshold. Therefore it would be required to train a separate model for each threshold and thus also increase the amount of parameters used in total.

Note that combination models of the type considered in this paper could also be constructed using artificial neural networks (ANN). For such models, there is no need to specify the explicit

parametric form between the underlying initial probabilistic forecasts and the event that is being predicted. Thus, ANN models may allow for more flexibility. Besides, it may also be possible to train a general ANN for the combination of forecasts, which can predict exceedance probabilities not only for one threshold, but for several thresholds simultaneously. In this case, the consistency of the calibrated probabilities has to be ensured, *i.e.*, the probabilities have to be smaller for increasing thresholds, see also Ben Bouallègue (2013).

The development of such ANN-based combination models for the prediction of several thresholds or a probability distribution will be the subject of a forthcoming paper.

Acknowledgments. The financial support by DWD (Deutscher Wetterdienst) for the project STOFOR through the extramural research program (EMF) is gratefully acknowledged. The authors also acknowledge support by the state of Baden-Württemberg through bwHPC.

APPENDIX A

Calibration

Using the same notation as before in this paper, let $f(P)$ be the self-calibrated version of a probabilistic forecast model P. It can be easily seen that $f(P)$ is calibrated in the sense of Eq. (2). Namely it holds that

$$\begin{aligned}\mathbb{E}(Y \mid f(P)) &= \mathbb{E}(Y \mid \mathbb{E}(Y \mid P)) \\ &= \mathbb{E}(Y \mid P) \\ &= f(P).\end{aligned}$$

This is a special case of the tower property of conditional expectation, which says that the identity $\mathbb{E}(X \mid \mathbb{E}(X \mid \mathscr{H})) = \mathbb{E}(X \mid \mathscr{H})$, holds for any random variable X and sub-σ-algebra $\mathscr{H}$ of $\mathscr{F}$. Note that the latter identity is sometimes called the Doob martingale property.

APPENDIX B

Sharpness

It turns out that $f(P)$ has the maximum variance compared to any other calibrated model $g(P)$ that is a function of P.

Indeed, let $g : [0,1] \to [0,1]$ be any deterministic function such that $g(P)$ is a well-defined random variable, which is calibrated, *i.e.*, $\mathbb{E}(Y \mid g(P)) = g(P)$. For brevity, we thereafter write f instead of $f(P)$, and g instead of $g(P)$. First, notice that

$$\mathrm{Var}(f) = \mathbb{E}(f^2) - q^2,$$

$$\mathrm{Var}(g) = \mathbb{E}(g^2) - q^2,$$

where $q = \mathbb{E}(Y)$. Then, it follows that

$$\mathrm{Var}(f) - \mathrm{Var}(g) = \mathbb{E}(f^2) - \mathbb{E}(g^2).$$

To show that $\mathbb{E}(f^2) - \mathbb{E}(g^2) \geq 0$, it suffices to observe that

$$\mathbb{E}((Y-g)^2) \geq \mathbb{E}((Y-f)^2)$$

as $f = \mathbb{E}(Y|P)$ is the orthogonal projection of Y on the L^2-space of square-integrable random variables. Besides,

$$\begin{aligned}
\mathbb{E}((Y-f)^2) &= \mathbb{E}(Y^2) - 2\mathbb{E}(Yf) + \mathbb{E}(f^2) \\
&= q - 2\mathbb{E}(\mathbb{E}(Yf \mid P)) + \mathbb{E}(f^2) \\
&= q - \mathbb{E}(f^2).
\end{aligned}$$

Note that the latter equality is straightforward because $\mathbb{E}(\mathbb{E}(Yf \mid P)) = \mathbb{E}(f\mathbb{E}(Y \mid P)) = \mathbb{E}(f^2)$ as $f(P)$ is $\sigma(P)$-measurable. With the same type of argument, one can show that $\mathbb{E}((Y-g)^2) =$

$q - \mathbb{E}(g^2)$. This gives that

$$q - \mathbb{E}(g^2) \geq q - \mathbb{E}(f^2)$$

and, thus, that $\mathbb{E}(f^2) \geq \mathbb{E}(g^2)$.

APPENDIX C

Limitation of f_{LT}

In this section a limitation of the combination model f_{LT} is shown, which can be resolved with additional coefficients that may be provided e.g. by the interaction terms in the combination model f_{LTI}. Consider the model f_{LT} with two initial forecasts P_1 and P_2:

$$f_{LT}(P_1, P_2) = \sigma\left(\sum_{i=1}^{2}\sum_{0=1}^{m} b_{ij}\phi_j(P_i)\right).$$

The triangular functions ϕ_j reach their maximum at $\frac{j}{m}$ with $\phi_j(\frac{j}{m}) = 1$ for each $j \in \{0, \ldots, m\}$. For the case where P_1 and P_2 take values in $\{0, \frac{1}{m}, \ldots, \frac{m-1}{m}, 1\}$ all triangular functions are zero, except for the two triangular functions, which take their maximum at $\frac{j_1}{m} = P_1$ and $\frac{j_2}{m} = P_2$. It then holds that

$$f_{LT}(P_1, P_2) = \sigma\left(\sum_{i=1}^{2}\sum_{0=1}^{m} b_{ij}\phi_j(P_i)\right) = \sigma\left(b_{1j_1}\phi_{j_1}(P_1) + b_{2j_2}\phi_{j_2}(P_2)\right) = \sigma\left(b_{1j_1} + b_{2j_2}\right). \quad \text{(C1)}$$

Now consider four points $(P_1', P_2'), (P_1'', P_2'), (P_1', P_2''), (P_1'', P_2'')$ with $P_1', P_1'', P_2', P_2'' \in \{0, \frac{1}{m}, \ldots, \frac{m-1}{m}, 1\}$, which form a rectangle similar to the crossing points of the four triangular functions in Fig. 3. For each of the four points, f_{LT} can be reduced as in Eq. C1:

$$f_{LT}(P_1', P_2') = \sigma(\ b_{1j_1'} \quad + b_{2j_2'} \quad),$$

$$f_{LT}(P_1'', P_2') = \sigma(\quad b_{1j_1''} + b_{2j_2'} \quad),$$

$$f_{LT}(P_1', P_2'') = \sigma(\ b_{1j_1'} \quad + \quad b_{2j_2''}),$$

$$f_{LT}(P_1'', P_2'') = \sigma(\quad b_{1j_1''} + \quad b_{2j_2''}).$$

These equations can be transformed into

$$\begin{aligned}
\sigma^{-1}(f_{LT}(P_1',P_2')) &= b_{1j_1'} \quad + b_{2j_2'},\\
\sigma^{-1}(f_{LT}(P_1'',P_2')) &= \quad b_{1j_1''} + b_{2j_2'},\\
\sigma^{-1}(f_{LT}(P_1',P_2'')) &= b_{1j_1'} \quad + \quad b_{2j_2''},\\
\sigma^{-1}(f_{LT}(P_1'',P_2'')) &= \quad b_{1j_1''} + \quad b_{2j_2''}.
\end{aligned}$$

Moreover, they can be written as a system of linear equations:

$$\left[\begin{array}{cccc|c}
1 & 0 & 1 & 0 & \sigma^{-1}(f_{LT}(P_1',P_2')) \\
0 & 1 & 1 & 0 & \sigma^{-1}(f_{LT}(P_1'',P_2')) \\
1 & 0 & 0 & 1 & \sigma^{-1}(f_{LT}(P_1',P_2'')) \\
0 & 1 & 0 & 1 & \sigma^{-1}(f_{LT}(P_1'',P_2''))
\end{array}\right].$$

Since the matrix is singular, it follows that in general there is no set of coefficients that would solve the system of linear equations and therefore the model f_{LT} can not satisfy the equations for all four points and will have to pick an approximate solution.

References

Ariely, D., W. Tung Au, R. H. Bender, D. V. Budescu, C. B. Dietz, H. Gu, T. S. Wallsten, and G. Zauberman, 2000: The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied*, **6 (2)**, 130.

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

Baars, J. A., and C. F. Mass, 2005: Performance of national weather service forecasts compared to operational, consensus, and weighted model output statistics. *Weather and Forecasting*, **20 (6)**, 1034–1047.

Baran, S., and S. Lerch, 2018: Combining predictive distributions for the statistical post-processing of ensemble forecasts. *International Journal of Forecasting*, **34 (3)**, 477–496.

Bassetti, F., R. Casarin, and F. Ravazzolo, 2018: Bayesian nonparametric calibration and combination of predictive distributions. *Journal of the American Statistical Association*, **113 (522)**, 675–685.

Ben Bouallègue, Z., 2013: Calibrated short-range ensemble precipitation forecasts using extended logistic regression with interaction terms. *Weather and Forecasting*, **28 (2)**, 515–524.

Bosart, L. F., 1975: Sunya experimental results in forecasting daily temperature and precipitation. *Monthly Weather Review*, **103 (11)**, 1013–1020.

Bottou, L., 2010: Large-scale machine learning with stochastic gradient descent. *Proceedings of COMPSTAT'2010*, Springer, 177–186.

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.

Chiu, S. N., D. Stoyan, W. S. Kendall, and J. Mecke, 2013: *Stochastic Geometry and its Applications*. J. Wiley & Sons.

Chollet, F., 2017: *Deep Learning with Python*. Manning Publications, URL https://www.ebook.de/de/product/28930398/francois_chollet_deep_learning_with_python.html.

Clemen, R. T., 1989: Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, **5 (4)**, 559–583.

Clemen, R. T., and R. L. Winkler, 1999: Combining probability distributions from experts in risk analysis. *Risk Analysis*, **19 (2)**, 187–203.

Genest, C., and K. J. McConway, 1990: Allocating the weights in the linear opinion pool. *Journal of Forecasting*, **9 (1)**, 53–73.

Gneiting, T., F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B*, **69 (2)**, 243–268.

Gneiting, T., R. Ranjan, and Coauthors, 2013: Combining predictive distributions. *Electronic Journal of Statistics*, **7**, 1747–1782.

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

Goodfellow, I., Y. Bengio, and A. Courville, 2016: *Deep Learning*. MIT Press.

Graham, J. R., 1996: Is a group of economists better than one? than none? *Journal of Business*, **69**, 193–232.

Gyakum, J. R., 1986: Experiments in temperature and precipitation forecasting for illinois. *Weather and Forecasting*, **1 (1)**, 77–88.

Hamill, T. M., R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Monthly Weather Review*, **136 (7)**, 2620–2632.

Hess, R., B. Kriesche, P. Schaumann, B. K. Reichert, and V. Schmidt, 2018: Area precipitation probabilities derived from point forecasts for operational weather and warning service applications. *Quarterly Journal of the Royal Meteorological Society*, **144 (717)**, 2392–2403.

Kober, K., G. 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.

631 Kriesche, B., R. Hess, B. K. Reichert, and V. Schmidt, 2015: A probabilistic approach to the
632 prediction of area weather events, applied to precipitation. *Spatial Statistics*, **12**, 15–30.

633 Kriesche, B., R. Hess, and V. Schmidt, 2017: A point process approach for spatial stochastic
634 modeling of thunderstorm cells. *Probability and Mathematical Statistics*, **37**, 471–496.

635 Murphy, A. H., and R. L. Winkler, 1977: Reliability of subjective probability forecasts of precipi-
636 tation and temperature. *Journal of the Royal Statistical Society: Series C*, **26 (1)**, 41–47.

637 Murphy, A. H., and R. L. Winkler, 1987: A general framework for forecast verification. *Monthly*
638 *Weather Review*, **115 (7)**, 1330–1338.

639 Pavlyshenko, B., 2018: Using stacking approaches for machine learning models. *2018 IEEE Sec-*
640 *ond International Conference on Data Stream Mining & Processing (DSMP)*, IEEE, 255–258.

641 Ranjan, R., and T. Gneiting, 2010: Combining probability forecasts. *Journal of the Royal Statisti-*
642 *cal Society: Series B*, **72 (1)**, 71–91.

643 Sanders, F., 1963: On subjective probability forecasting. *Journal of Applied Meteorology*, **2 (2)**,
644 191–201.

645 Tashman, L. J., 2000: Out-of-sample tests of forecasting accuracy: an analysis and review. *Inter-*
646 *national Journal of Forecasting*, **16 (4)**, 437–450.

647 Theis, S., A. Hense, and U. Damrath, 2005: Probabilistic precipitation forecasts from a determin-
648 istic model: A pragmatic approach. *Meteorological Applications*, **12 (3)**, 257–268.

649 Vislocky, R. L., and J. M. Fritsch, 1995: Improved model output and statistics through model
650 consensus. *Bulletin of the American Meteorological Society*, **76 (7)**, 1157–1164.

651 Weigl, E., and T. Winterrath, 2010: Radargestützte niederschlagsanalyse und-vorhersage (radolan,
652 radvor-op). *Promet*, **35**, 78–86.

653 Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*, Vol. 91. Academic Press.

654 Wilks, D. S., 2009: Extending logistic regression to provide full-probability-distribution mos fore-
655 casts. *Meteorological Applications*, **16 (3)**, 361–368.

656 Winkler, R., A. Murphy, and R. Katz, 1977: The consensus of subjective probability forecasts:
657 Are two, three,..., heads better than one. *5th Conference on Probability and Statistics*, 57–62.

658 Winkler, R. L., and R. M. Poses, 1993: Evaluating and combining physicians' probabilities of
659 survival in an intensive care unit. *Management Science*, **39 (12)**, 1526–1543.

660 Winterrath, T., and W. Rosenow, 2007: A new module for the tracking of radar-derived precipita-
661 tion with model-derived winds. *Advances in Geosciences*, **10**, 77–83.

662 Winterrath, T., W. Rosenow, and E. Weigl, 2012: On the dwd quantitative precipitation analysis
663 and nowcasting system for real-time application in german flood risk management. *Weather*
664 *Radar and Hydrology, IAHS Publ*, **351**, 323–329.

LIST OF FIGURES

Fig. 1. Reliability diagrams of the considered (initial and combined) probabilistic forecasts for all considered locations and for three lead times (+1h, +3h, +6h). The superimposed bar plots show the empirical distribution of the forecast values over the unit interval. The x-axis represents the forecast probability and the y-axis the observed relative frequency. The upper and lower ends of the grey band correspond to the 95% and 5% quantiles of the reliability diagrams for single locations to quantify the calibration of each forecast model at single locations. 37

Fig. 2. Six triangular functions for $m = 5$. 38

Fig. 3. Exemplary effect of single triangular functions on the output of the combination model. For these plots, most of the coefficients are set to zero, except for a few, to highlight the shape and the interplay between single triangular functions. The color indicates the predicted probability by the combination model for pairs of initial forecasts. Since it holds for the sigmoid function that $\sigma(0) = 0.5$, the area unaffected by the triangular functions with non-zero coefficients is green. **Top left:** Two triangular functions of Ensemble-MOS with coefficients -1 and 2. **Top right:** Three triangular functions of γ_1 with coefficients -2, 1 and 3. **Bottom left:** Interplay between two triangular functions of Ensemble-MOS and RadVOR each. **Bottom right:** Interplay between triangular functions from the bottom left and top right plot. 39

Fig. 4. Comparison of validation scores for different combinations of hyperparameters for the lead times $+1h$, $+3$ and $+6h$. The hyperparameters used for the results presented in this paper ($\eta_0 = 0.0005$, $n = 11$) are marked with a dot. Note that the color bars are not linear and that the absolute value of the bias is shown. 40

Fig. 5. **Upper plots:** Precipitation probabilities predicted by the fitted combination model f_{LTI} for the months of June and July 2016 for pairs of initial forecasts. **Lower plots:** Average observed probability of precipitation for pairs of RadVOR/Ensemble-MOS forecasts for the months of June and July 2016. Initial forecast pairs, which occur less than 50 times, are left blank. 41

Fig. 6. Evolution of bias, Brier skill score, reliability and sharpness of the considered (initial and combined) probabilistic forecasts with respect to various lead times. The box plot diagrams show the behavior of the daily-averages of the scores and the continuous lines the averages over all locations and time periods. The x-axis represents the lead times of the forecasts and the y-axis the scores values. 42

Fig. 7. Distribution of the time-dependent parameters a_{ij} and b_{ij} for each triangular function of the model f_{LTI} for June (red) and July (blue) for lead times +1h, +3h and +6h. 43

Fig. 8. Average bias for single locations for the lead times from $+1h$ to $+6h$. Locations with a bias above 0.05 are shown in violet. 44

Fig. 9. Average Brier skill score for single locations for the lead times from $+1h$ to $+6h$. Locations with a brier skill score above 0.7 are shown in grey. Locations with a brier skill score below 0 are shown in violet. 45

Fig. 10. Average reliability for single locations for the lead times from $+1h$ to $+6h$. Locations with a reliability above 0.05 are shown in violet. Locations with a reliability below 0.001 are shown in grey. 46

Fig. 11. A case study for the combination of point probabilities for a single hour (14th July 2016 from 10:00 to 11:00) for the lead times from $+6h$ down to $+1h$. 47

Fig. 12. Forecast scores for area probabilities based on RadVOR, Ensemble-MOS and their combination. In case of Method 1, the point probabilities given by RadVOR and Ensemble-MOS, respectively, are first combined and then converted into area probabilities. In case of Method 2, both sets of point probabilities are first converted into area probabilities and then combined. The x-axis represents the lead times of the forecasts and the y-axis the scores values. 48

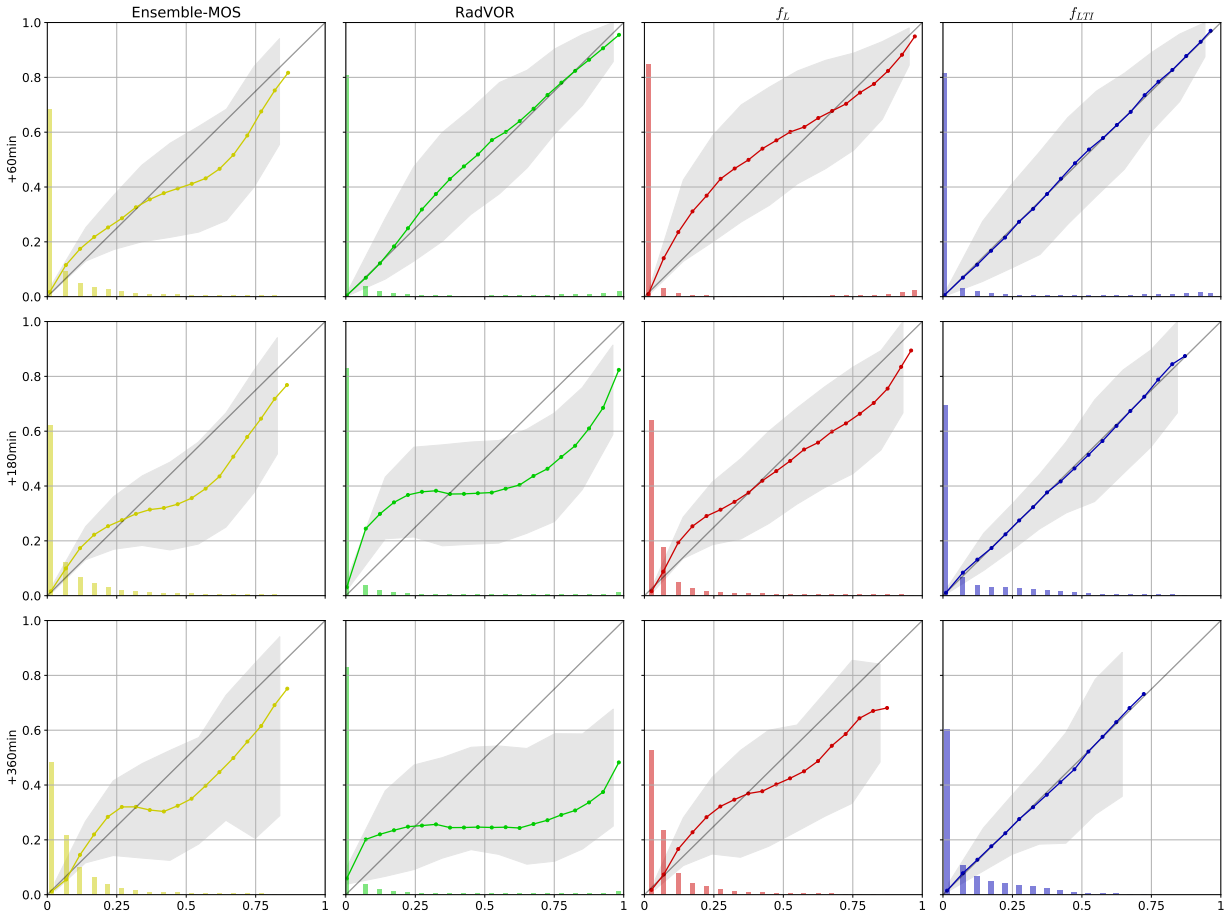

FIG. 1: Reliability diagrams of the considered (initial and combined) probabilistic forecasts for all considered locations and for three lead times (+1h, +3h, +6h). The superimposed bar plots show the empirical distribution of the forecast values over the unit interval. The x-axis represents the forecast probability and the y-axis the observed relative frequency. The upper and lower ends of the grey band correspond to the 95% and 5% quantiles of the reliability diagrams for single locations to quantify the calibration of each forecast model at single locations.

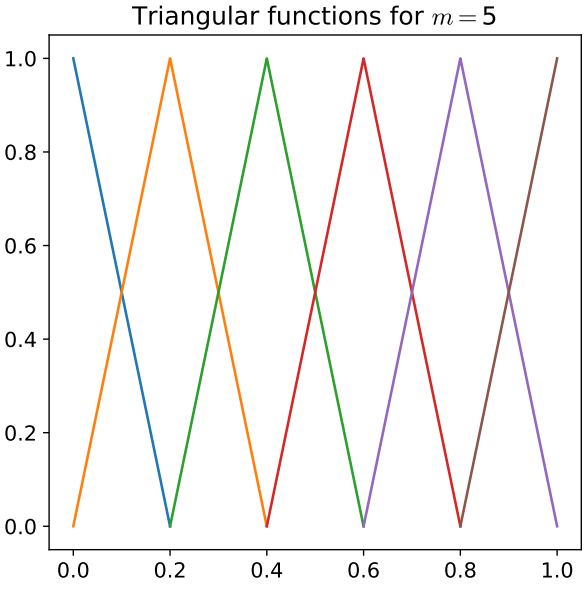

FIG. 2: Six triangular functions for $m = 5$.

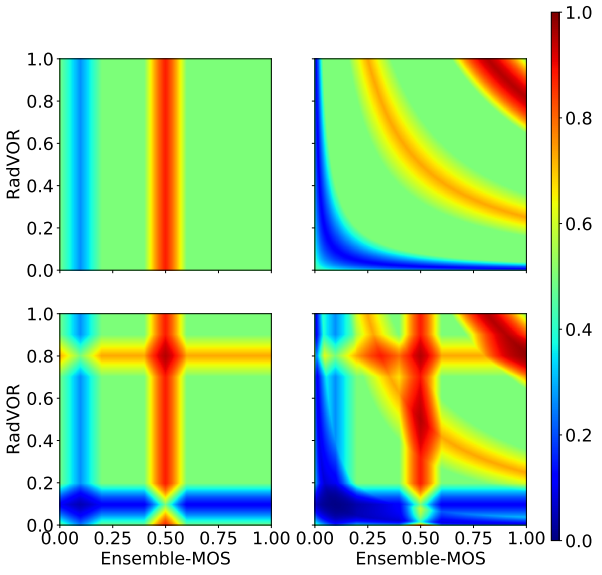

FIG. 3: Exemplary effect of single triangular functions on the output of the combination model. For these plots, most of the coefficients are set to zero, except for a few, to highlight the shape and the interplay between single triangular functions. The color indicates the predicted probability by the combination model for pairs of initial forecasts. Since it holds for the sigmoid function that $\sigma(0) = 0.5$, the area unaffected by the triangular functions with non-zero coefficients is green.
Top left: Two triangular functions of Ensemble-MOS with coefficients -1 and 2.
Top right: Three triangular functions of γ_1 with coefficients -2, 1 and 3.
Bottom left: Interplay between two triangular functions of Ensemble-MOS and RadVOR each.
Bottom right: Interplay between triangular functions from the bottom left and top right plot.

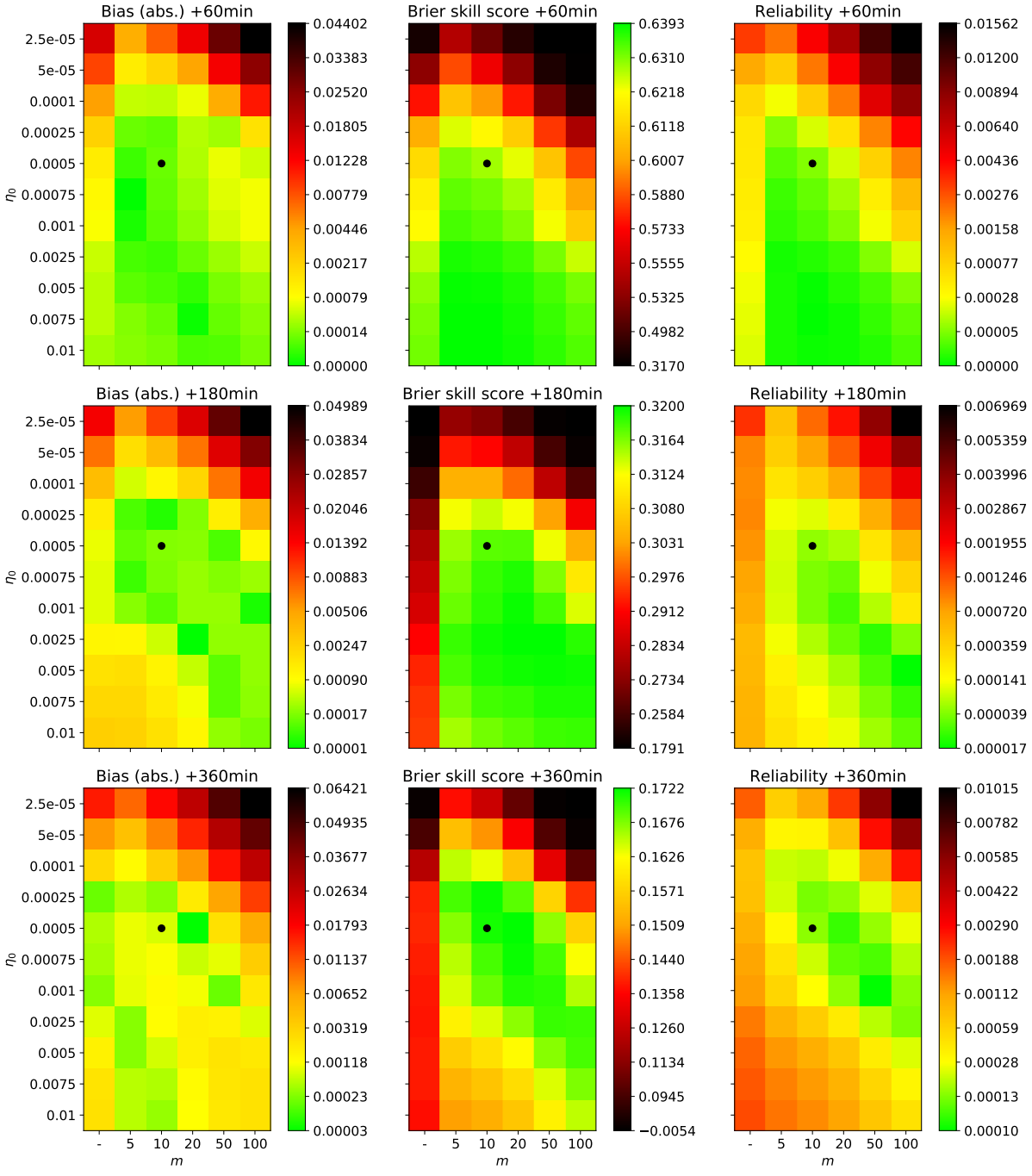

FIG. 4: Comparison of validation scores for different combinations of hyperparameters for the lead times $+1h$, $+3$ and $+6h$. The hyperparameters used for the results presented in this paper ($\eta_0 = 0.0005$, $n = 11$) are marked with a dot. Note that the color bars are not linear and that the absolute value of the bias is shown.

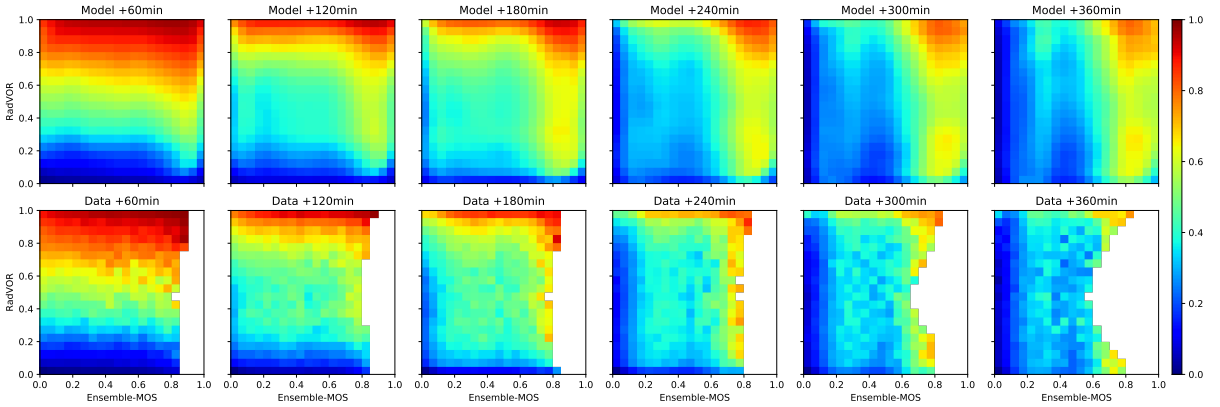

FIG. 5: **Upper plots:** Precipitation probabilities predicted by the fitted combination model f_{LTI} for the months of June and July 2016 for pairs of initial forecasts.
Lower plots: Average observed probability of precipitation for pairs of RadVOR/Ensemble-MOS forecasts for the months of June and July 2016. Initial forecast pairs, which occur less than 50 times, are left blank.

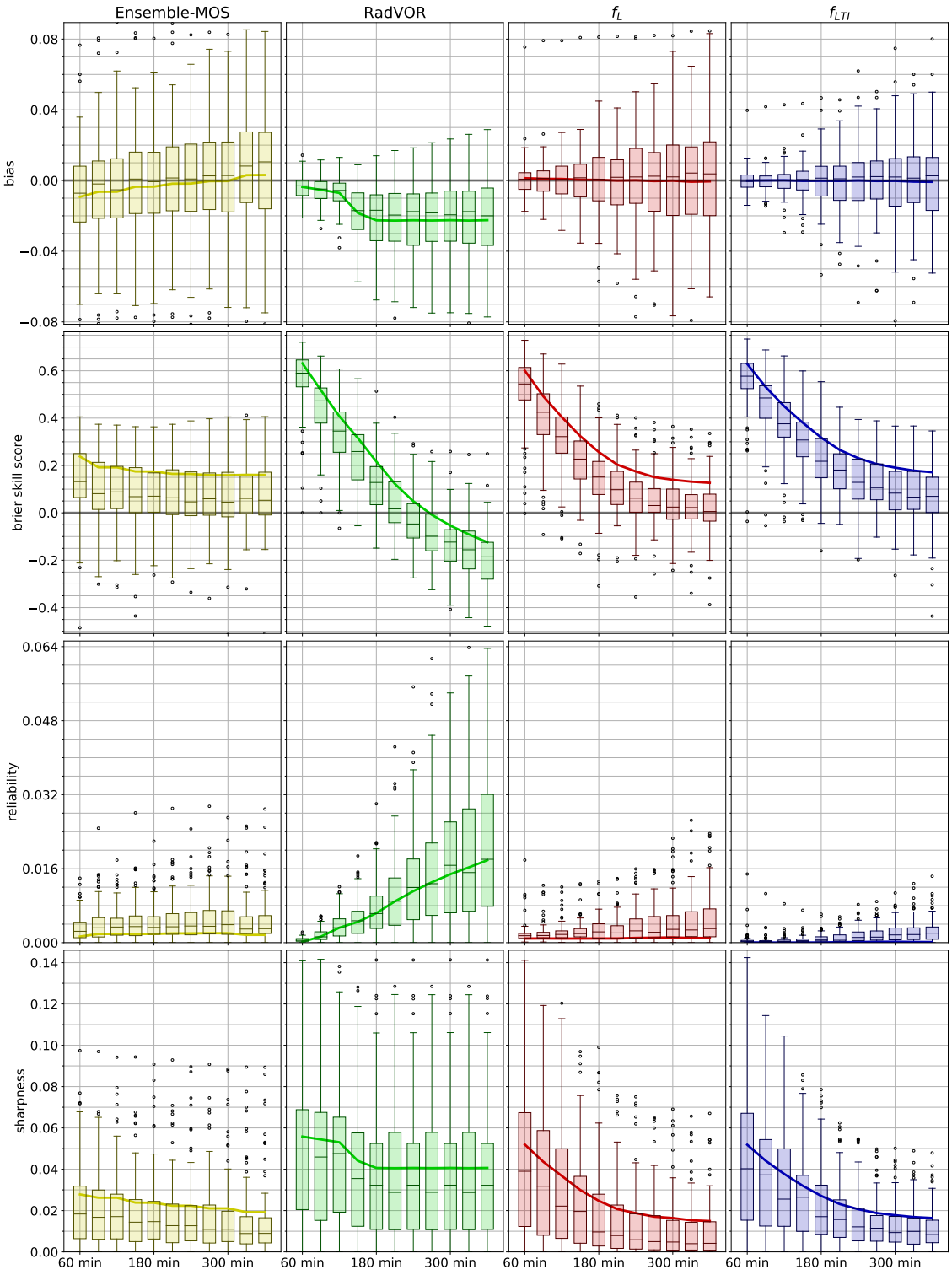

FIG. 6: Evolution of bias, Brier skill score, reliability and sharpness of the considered (initial and combined) probabilistic forecasts with respect to various lead times. The box plot diagrams show the behavior of the daily-averages of the scores and the continuous lines the averages over all locations and time periods. The x-axis represents the lead times of the forecasts and the y-axis the scores values.

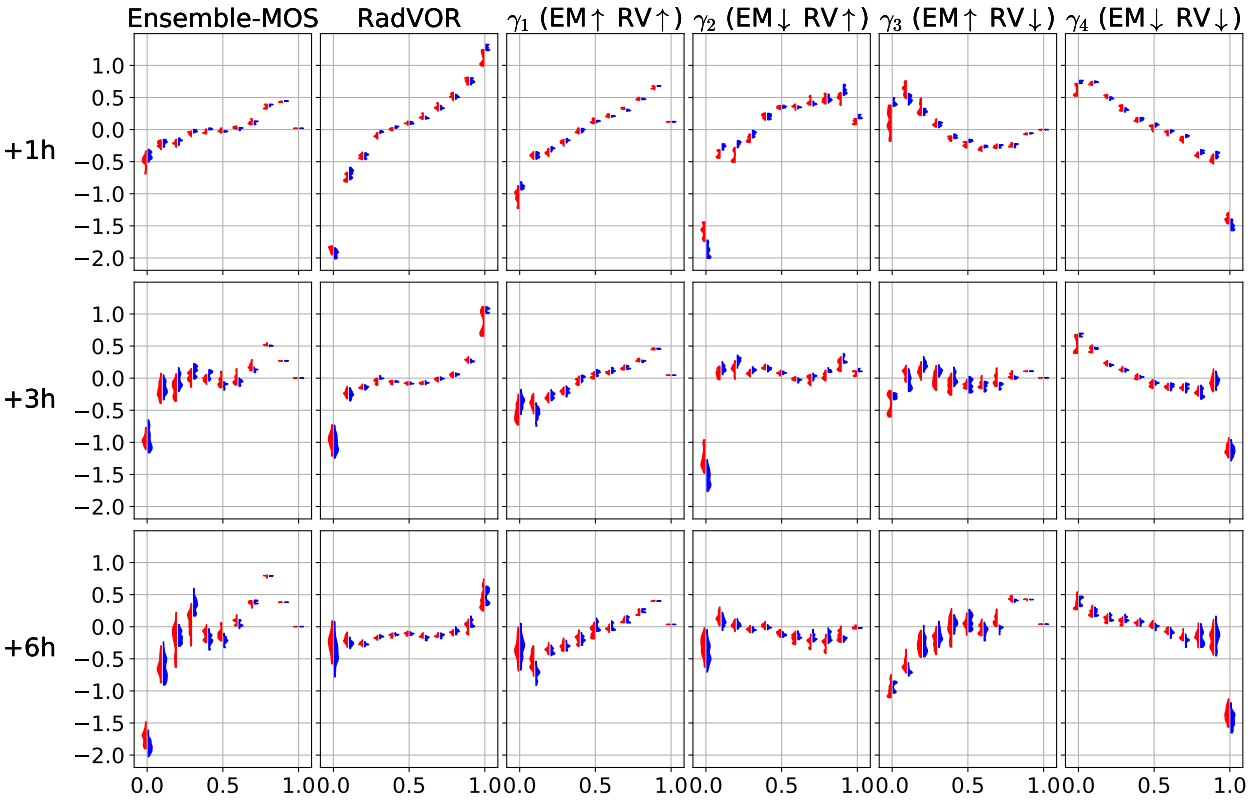

FIG. 7: Distribution of the time-dependent parameters a_{ij} and b_{ij} for each triangular function of the model f_{LTI} for June (red) and July (blue) for lead times +1h, +3h and +6h.

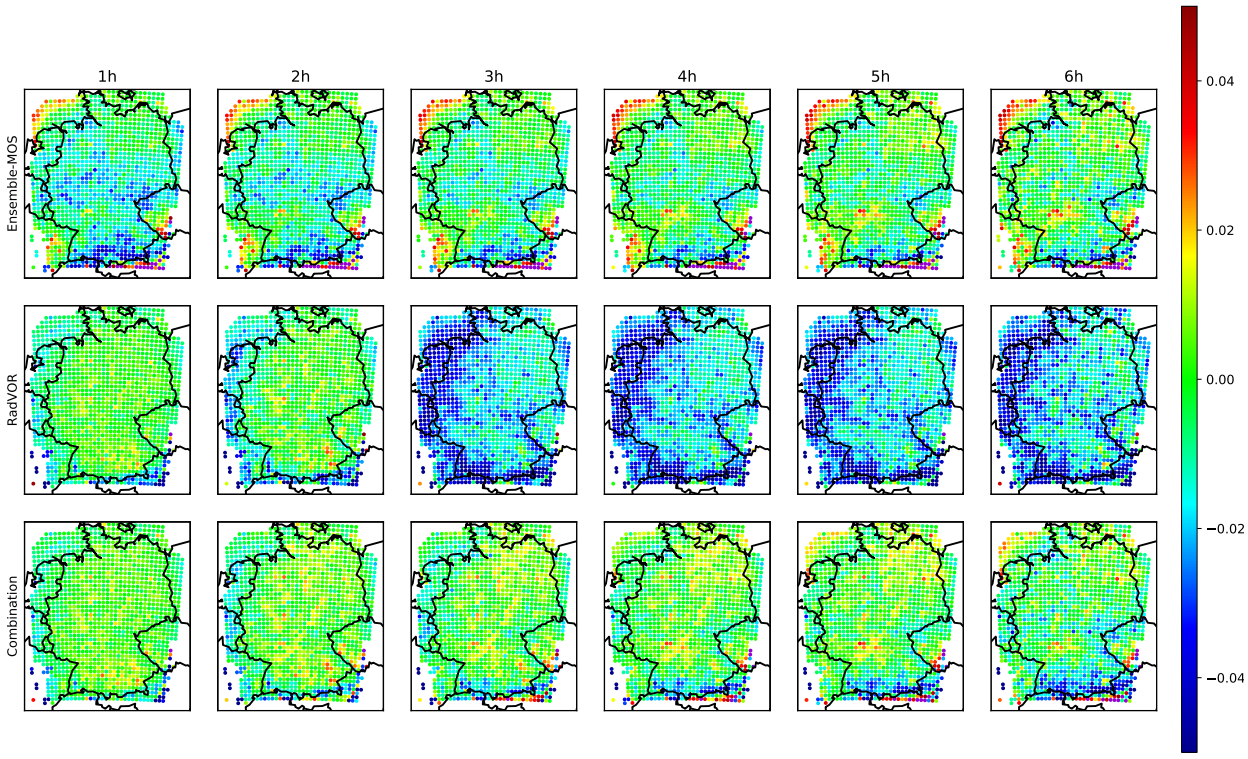

FIG. 8: Average bias for single locations for the lead times from $+1h$ to $+6h$. Locations with a bias above 0.05 are shown in violet.

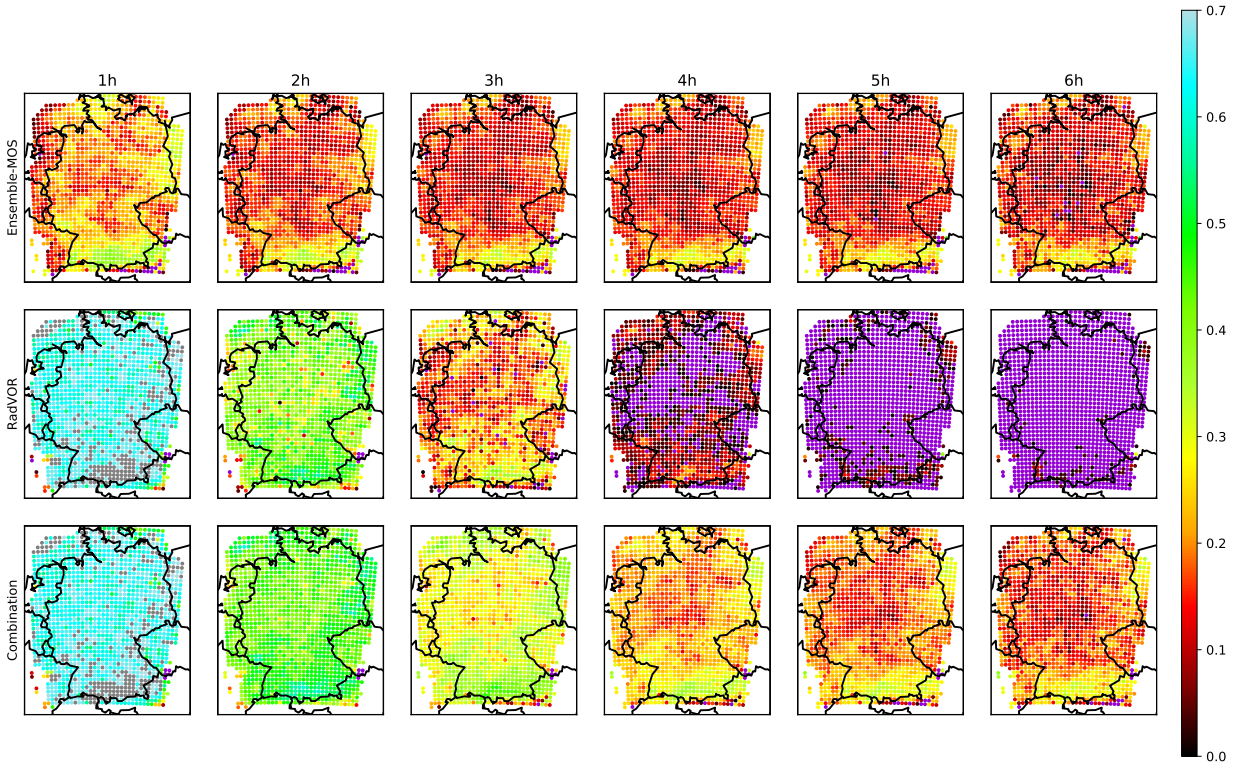

FIG. 9: Average Brier skill score for single locations for the lead times from $+1h$ to $+6h$. Locations with a brier skill score above 0.7 are shown in grey. Locations with a brier skill score below 0 are shown in violet.

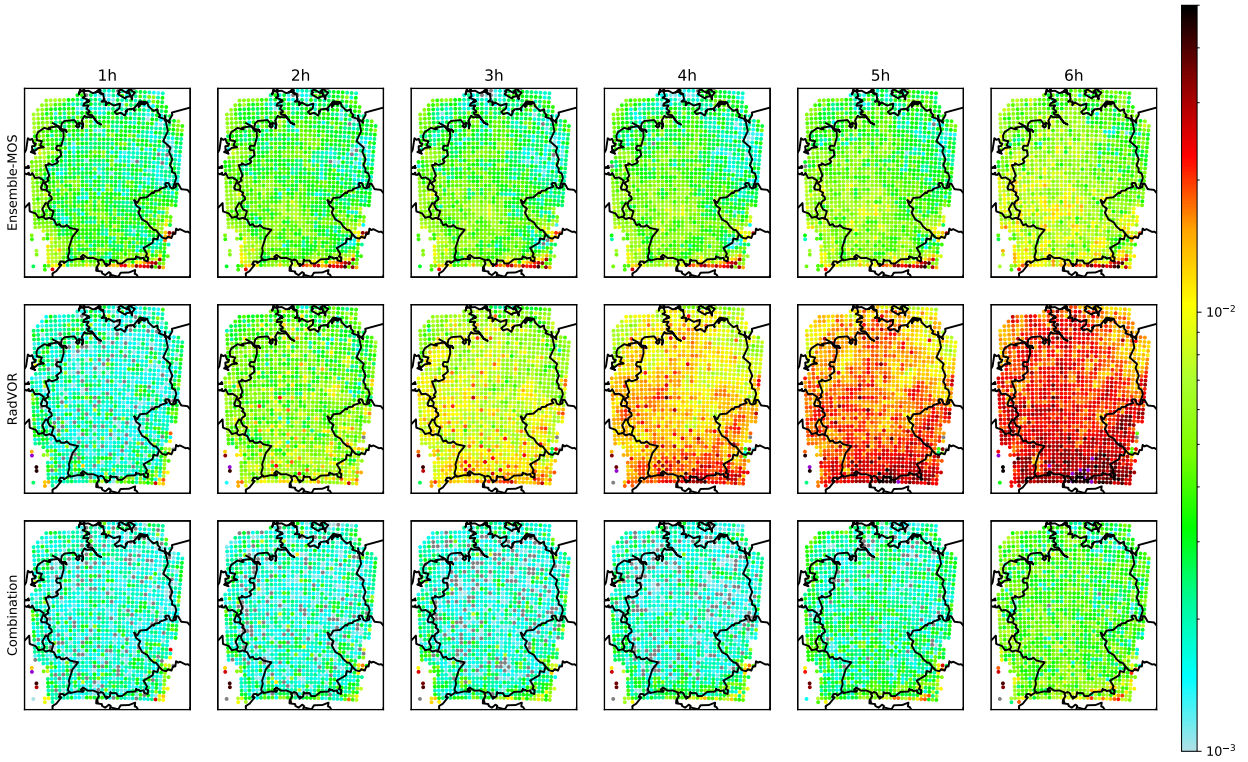

FIG. 10: Average reliability for single locations for the lead times from $+1h$ to $+6h$. Locations with a reliability above 0.05 are shown in violet. Locations with a reliability below 0.001 are shown in grey.

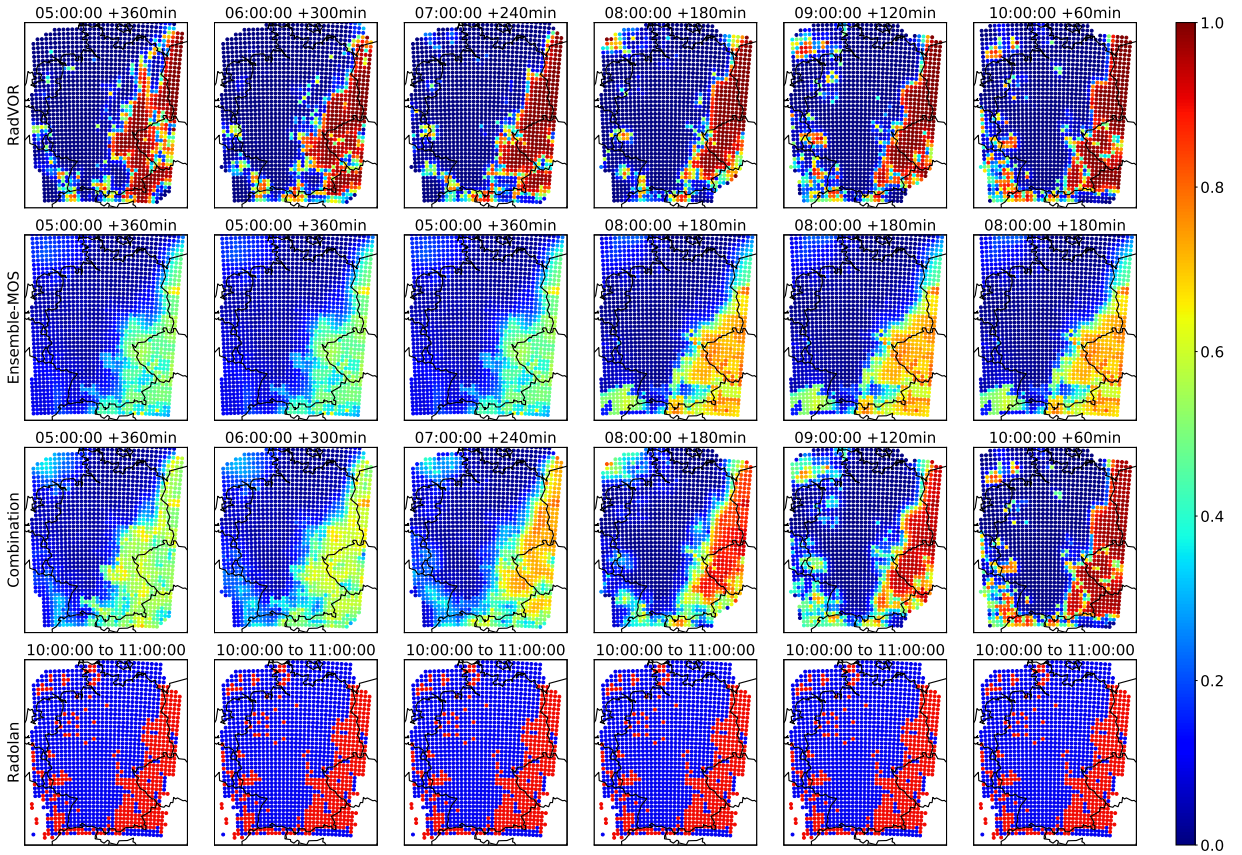

FIG. 11: A case study for the combination of point probabilities for a single hour (14th July 2016 from 10:00 to 11:00) for the lead times from $+6h$ down to $+1h$.

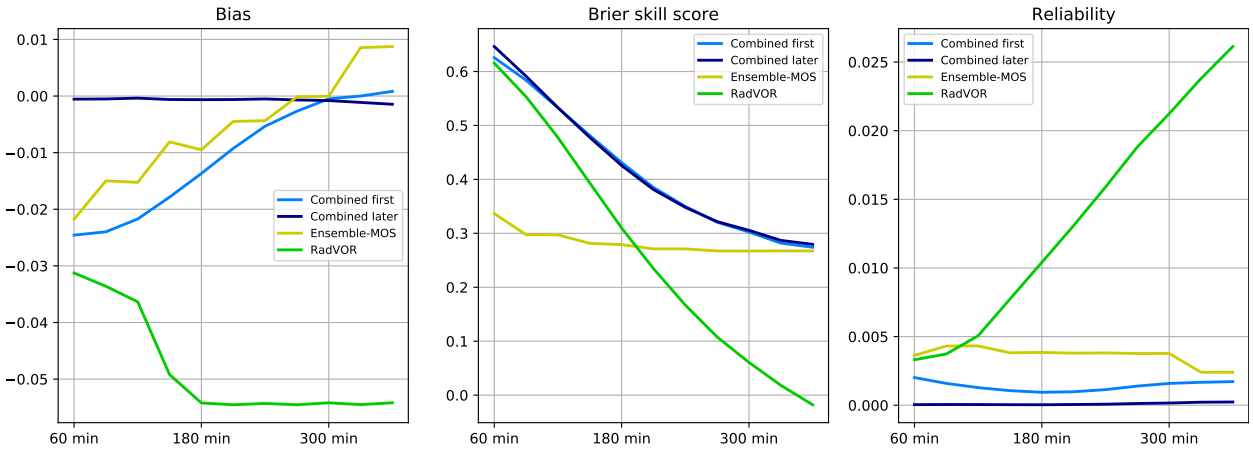

FIG. 12: Forecast scores for area probabilities based on RadVOR, Ensemble-MOS and their combination. In case of Method 1, the point probabilities given by RadVOR and Ensemble-MOS, respectively, are first combined and then converted into area probabilities. In case of Method 2, both sets of point probabilities are first converted into area probabilities and then combined. The x-axis represents the lead times of the forecasts and the y-axis the scores values.