Neighborhood-Based Ensemble Evaluation Using the CRPS

JOEL STEIN
A and FABIEN STOOP
A

Météo-France, Toulouse, France

(Manuscript received 25 August 2021, in final form 1 April 2022)

ABSTRACT: The neighborhood-based ensemble evaluation using the continuous ranked probability score is based on the pooling of the cumulative density function (CDF) for all the points inside a neighborhood. This methodology can be applied to the forecast CDF for measuring the predictive input of neighboring points in the center of the neighborhood. It can also be applied at the same time to forecast CDF and observed CDF so as to quantify the quality of the pooled ensemble forecast at the scale of the neighborhood. Fair versions of these two neighborhood scores are also defined in order to reduce their dependencies on the size of ensemble forecasts. The borderline case of deterministic forecasts is also explored so as to be able to compare them with ensemble forecasts. The information of these new scores is analyzed on idealized and real cases of rain accumulated during 3 h and of 2-m temperature forecast by four deterministic and probabilistic forecasting systems operational at Météo-France.

KEYWORDS: Forecast verification/skill; Operational forecasting; Probability forecasts/models/distribution

1. Introduction

The verification of meteorological forecasts is an unavoidable subject if one wants to be able to improve a forecast system. Numerous methods exist (Jolliffe and Stephenson 2011), each one possessing its application area among a wide range of forecast types: deterministic or probabilistic, forecasts of binary events, multi-categorical, forecasts of continuous variables. In most centers of numerical weather prediction, the deterministic forecasts are evaluated by the calculation of scores such as the root of the mean square error, the mean Absolute Error (AE below), and the correlation of anomalies. In addition to these overall quality measures, evaluations of binary events, such as the presence of fog, rain, snow, and wind stronger than 100 km h\(^{-1}\), can be included by using specific scores deduced from contingency tables which class the forecasts into four categories: hits, false alarms, misses, and correct rejections. A desirable property for this type of scores is the equitability, which guarantees that the constant forecasts and the random forecasts all have the same score (Jolliffe and Stephenson 2011). We invite the reader to consult the reference books of Jolliffe and Stephenson (2011) and of Wilks (2011) for a more thorough presentation of these scores.

A new stage has been reached through the introduction of neighborhood in the verification of binary cases by Roberts and Lean (2008). In the case of a forecast of a successfully observed but slightly staggered event, we will count in the case of a strict control a miss at the observation point and a false alarm at the expected location. This counting of a double error is called the double penalty. The double penalty particularly penalizes forecasts of high-resolution models which often explicitly represent difficult-to-predict small scale phenomenon with small location errors. The larger scale models tend to spread this type of phenomenon across large zones and underestimate their intensity. The introduction of a neighborhood consists of creating a tolerance on the spatial error of the forecast and thus allowing the reduction of the impact on the scores of this double penalty. Ebert (2008) makes a classification of different verification methods calling on the neighborhood according to whether the observations are only considered in the center of the neighborhood (single observation or “so” thereafter) or across the whole of the neighborhood (neighborhood observation or “no” thereafter). Roberts and Lean (2008) use the neighborhood frequency of events defined as the frequency of the event in a neighborhood. They compare the forecast neighborhood frequency to the observed neighborhood frequency of events in the same neighborhood. They consider an average quadratic error to quantify the difference between the fields of neighborhood frequencies. Several normalizations have been introduced to transform the absolute score into a relative score. Roberts and Lean (2008) use a linearly decorrelated forecast of the observation leading to the fractions skill score (FSS) while Amodei and Stein (2009) use the same score with a normalization via persistence forecasting providing therefore a neighborhood Brier skill score. These indicators are used operationally at the Met Office (Mittermaier et al. 2013) and at Météo-France (Amodei et al. 2015). Different trials to include the notion of neighborhood directly in the contingency tables have been reviewed by Schwartz (2017) with more or less symmetrical treatments of forecasts and observations on the local and neighborhood scales. Stein and Stoop (2019) propose an original method of error compensation inside the neighborhood between false alarms and misses which allows the building of a contingency table that takes into account the neighborhood and from which one can calculate all the associated contingency scores.

© 2022 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

Denotes content that is immediately available upon publication as open access.

DOI: 10.1175/MWR-D-21-0224.1
Probabilistic forecasts have been around for decades but the growing use of ensemble forecasts (using perturbed initial conditions and perturbated versions of the numerical model) means that probabilistic forecasts are now more common and in routine use than in earlier decades. One can cite as traditional measures of their quality: for the probabilistic forecasts of continuous outcome variables, the continuous ranked probability score (CPRS) below (Matheson and Winkler 1976; Hersbach 2000) and for binary events, the Brier score; the ROC curve (Mason 1982; Stanski et al. 1989) and the economic value (Richardson 2000). The part devoted to the verification of ensemble forecasts expands in each new edition of reference works concerning the verification of forecasts (Jolliffe and Stephenson 2011; Wilks 2011) in discussing the different properties of these scores such as that of propriety (the fact of being proper), which ensures that such a score cannot be faulted by a forecast strategy other than to improve forecasting (Gneiting and Raftery 2007). New promising scores have been recently proposed by Ben Bouallégue et al. (2018) such as the forecast skill card and the diagonal score, which accumulates the equitability and propriety properties. Fricker et al. (2013) and Ferro (2014) have introduced the notion of fair scores so as to better reward the ensemble forecasts, which are issued from the same probability distribution as the observations. Ferro (2014) proposes fair versions for the Brier score and for the CPRS. Zamo and Naveau (2018) have verified for an idealized case that only the fair estimator of the CPRS was nonbiased for ensemble forecasts of small size.

The inclusion of neighborhoods for ensemble forecasts has been realized in different ways (Schwartz and Sobash 2017). Schwartz et al. (2010) use forecasts with neighboring points from the point of interest, as well as possible realizations with the central point of the neighborhood so as to calculate an average probability on the neighborhood. The field of this averaged probability is smoother than the field of the probability obtained with a single forecast at the central point for each member. This leads to increase the effective size of the ensemble which becomes equal to the initial size of the ensemble multiplied by the number of neighboring points. Ben Bouallégue and Theis (2014) propose to replace the local probability of exceeding a given threshold by a higher probability obtained by replacing the local forecasts of each member by the maximum in the neighborhood of the forecasts of each member. In these two cases, we are talking about transformations of the forecasts by post-processings destined to improve the forecasts or making products more reactive. In the same way, Taillardat and Mestre (2020) propose the improvement of the ensemble rain forecasts of a model with kilometer resolution using a statistical calibration based on a minimization of the CRPS evaluated in a neighborhood which is larger than the output grid of the forecasts by quantile regression techniques, using random forests and a semiparametric extension for the tail of extreme precipitation. Mittermaier (2014) proposes a framework to evaluate deterministic forecasts through probabilistic metrics applied to neighborhoods of forecast grid points. This framework referred to as the High Resolution Assessment Framework (HiRA) is then used to compare deterministic and ensemble forecasts for a range of forecast neighborhood sizes centered on surface site locations (Mittermaier and Csima 2017). A recent article of Zhao et al. (2021) presents a CRPS-based spatial technique named SCRPS, which applies to binary events. Like Roberts and Lean (2008), SCRPS is based on the neighborhood frequencies of the event in the neighborhood I_o for the observations and I_m, m = 1, …, M for the different members of the ensemble forecast, where M is the number of members. SCRPS is defined by

\[
\text{SCRPS} = \mathbb{E}_{I_o}(|I_m - I_o|) - \frac{1}{2} \mathbb{E}_{I_{o}I_m}(|I_m - I_m'|),
\]

where \( \mathbb{E}_{X,X'}(\mathbb{E}_X) \) is the expectation according to the joint law of random variables \( X \) and \( X' \) (\( X \)).

The aim of this article is to present how to integrate the notion of neighborhood in the verification of ensemble forecasts by pooling the forecasts or the observations for all the points of the neighborhood. The comparison of the forecast or observed pooled cumulative density functions (CDF) is performed in section 2 using either the CRPS or the score divergence associated to the CRPS (Thorarinsdottir et al. 2013). These scores are used for the idealized cases in section 3. Section 4 will present a verification of the quantitative precipitation and of 2-m temperature forecast by two ensemble forecasts at high and low resolutions as well as by two deterministic forecasts with higher resolution than that of members of the ensemble associated forecasts. Section 5 contains a discussion of the results and section 6 presents the conclusions.

2. Neighborhood pooling strategy to evaluate ensemble forecasts

a. Verification against single observations

Let us designate by \( x \) the forecast value of the outcome variable (rain, temperature, …), \( y \) its observed value, and \( F \) the CDF. The CRPS (Matheson and Winkler 1976; Hersbach 2000) measures the difference between the probabilistic forecast given by its CDF \( F \) and the observation \( y \) at a given point by the following formula:

\[
\text{CRPS}(F,y) = \int_{-\infty}^{+\infty} [F(x) - H(x - y)]^2 \, dx,
\]

where \( H(x) = 1 \) if \( x \geq 0 \) and 0 otherwise, is the Heaviside function. The minimum value of CRPS is obtained in the case of ensemble forecasts when all the \( M \) members forecast values \( x(m) \) for \( m = 1, 2, \ldots, M \) are equal to the observation \( y \). This score is negatively oriented. It takes into account the whole range of values of the outcome variable under consideration unlike the case of the Brier score designed for categorical outcomes (Jolliffe and Stephenson 2011). CRPS has the same unit as the outcome variable. These qualities render this score very attractive to evaluate the ensembles of forecasts. The integral in the right-hand side of the Eq. (2) can be numerically evaluated by decomposing it over the different intervals where the function to integrate is constant (Hersbach 2000). The observations at the \( N \) different points in the verification
area and the \( N_t \) different dates of the time sample are supposed independent draws from the empirical distribution \( G \) of the observations. The average score \( \text{CRPS}(F, G) \) is then given by

\[
\text{CRPS}(F, G) = \frac{1}{N_t N_M} \sum_{i=1}^{N_t} \sum_{m=1}^{M} \text{CRPS}(F, y_{x,i}),
\]

(3)

where \( y_{x,i} \) refers to the observation value at the spatial position \( s \) at time \( t \).

It has been demonstrated that the CRPS is a strictly proper score (Gneiting and Raftery 2007) when \( F \) has a finite first moment. A supplementary advantage of the CRPS is that it allows a coherent comparison between deterministic and probabilistic forecasts (Gneiting and Raftery 2007). Indeed, it is easy to show that in the deterministic case, the average CRPS degenerates to the mean AE defined by

\[
\text{AE} = \frac{1}{N_t N_M} \sum_{i=1}^{N_t} \sum_{m=1}^{M} |x_{i,m} - y_{x,i}|,
\]

(4)

where \( x_{i,m} \) refers to the single forecast value at the spatial position \( s \) at time \( t \).

Baringhaus and Franz (2004) and Székely and Rizzo (2005) have proposed a more general formulation allowing the evaluation of this type of distance between two CDFs \( F \) and \( G \):

\[
\int_{-\infty}^{\infty} |F(x) - G(x)|^2 \, dx = \mathbb{E}_{X,Y}(|X - Y|) - \frac{1}{2} [\mathbb{E}_{X,X'}(|X - X'|) + \mathbb{E}_{Y,Y'}(|Y - Y'|)],
\]

(5)

where \( X, X' (Y, Y') \) are independent copies of a random variable with a CDF given by \( F (G) \) having a finite first moment. Gneiting and Raftery (2007) use this formula in order to obtain the energy formulation of the CRPS in the case of a point measure for the observation \( y \):

\[
\text{CRPS}(F, y) = \mathbb{E}_{X}(|X - y|) - \frac{1}{2} [\mathbb{E}_{X,X'}(|X - X'|)].
\]

(6)

This makes more explicit that the CRPS generalizes the AE so as to reduce if the forecast density function is also a point measure for a deterministic forecast. Zamo and Naveau (2018) have verified for an idealized case that both computations provided by Eqs. (2) and (6) give the same result.

The introduction of the neighborhood is realized as in Schwartz et al. (2010), the \( N_n \) points of the neighborhood form indiscernible realizations of those of the central point for the forecasts. We consider therefore an ensemble forecast with \( N_n M \) members. The location inside the neighborhood is no longer important as all the points are equivalent. In other words, the forecasts realized by the different members at any point of the neighborhood are all considered coming from the same CDF \( F \). The CDF \( F \) is pooled for all the points in the neighborhood to provide a pooled CDF \( F_n \) which can be explicitly written as

\[
F_n(x) = \frac{1}{N_n M} \sum_{i=1}^{N_n} \sum_{m=1}^{M} H[x - x(i,m)],
\]

(7)

where \( x(i, m) \) represents the forecast value of member \( m \) at the point \( i \) of the neighborhood.

However, we compare the neighborhood forecast CDF \( F_n \) built with \( N_n M \) members to the observed value \( y \) in the center of the neighborhood so as to finally obtain a CRPS expression including the neighborhood only for the forecasts. In a nutshell, we perform an implicit pooling postprocessing step before calculating the CRPS. This expression of the pooled ensemble forecasts is named CRPSuso: “u” so as to remind us that it is unfair (Ferro 2014) and “so” that it only uses the single central observation:

\[
\text{CRPSuso}(F, y) = \text{CRPS}(F_n, y).
\]

(8)

Its local evaluation is provided by

\[
\text{CRPSuso}(F, y) = \frac{1}{N_n M} \sum_{i=1}^{N_n} \sum_{m=1}^{M} |x(i, m) - y| - \frac{1}{2} \left[ \frac{1}{N_n M^2} \sum_{i=1}^{N_n} \sum_{m=1}^{M} \sum_{j=1}^{N_n} \sum_{n=1}^{M} |x(i,m) - x(j,n)| \right].
\]

(9)

It should be clear that the CRPSuso evaluates the possible improvement brought by using a neighborhood as a postprocessing of the original ensemble forecast through the computation of the original proper score CRPS given by Eqs. (2) or (6). The CRPSuso is then averaged in the same way as the CRPS by Eq. (3). We can see that the energy formula can easily be applied to calculate the CRPS in the case described by Schwartz et al. (2010) or Ben Bouallégue and Theis (2014). This score allows the quantification of the improvement brought by the neighborhood pooling before calculating local probabilities as Mittermaier and Csima (2017) do in the HiRa framework.

The borderline case of a deterministic forecast is obtained for the CRPSuso by considering an ensemble with one member. The CDF \( F_n \) is given in this case by Eq. (7) with \( M = 1 \) and the CRPSuso becomes

\[
\text{CRPSuso}_{\text{det}}(F, y) = \frac{1}{N_n} \sum_{i=1}^{N_n} |x(i) - y| - \frac{1}{2} \left[ \frac{1}{N_n^2} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} |x(i) - x(j)| \right].
\]

(10)

The first term corresponds to a AE averaged on the neighborhood and the second corresponds to the dispersion of expected values in the neighborhood. In fact, the CRPSuso_{\text{det}} measures the quality of the replacement of the deterministic forecast at the central point with ensemble forecasts of which the \( N_n \) members are provided by the deterministic forecasts at the \( N_n \) points of the neighborhood, in the same way as in the HiRa framework (Mittermaier 2014).
b. Verification against neighborhood observations

To use neighborhood observations, we have to switch from a comparison between forecast CDF and point observations represented by a Heaviside function to the comparison of two CDFs. This is the purpose of Eq. (5): its lhs corresponds to the definition of the divergence function \(d(F, G)\) of the Cramer–Von Mises type and its rhs is the more general energy distance, a metric that measures the distance between the distributions of random vectors (Rizzo and Székely 2016). They are equal for real-valued random variables, which is our case. Both distances have the right properties to rank different forecasting systems against the empirical distribution of the observations (Rizzo and Székely 2016). Thorarinsdottir et al. (2013) call the divergence function associated to the CRPS, the integrated quadratic distance \(d_{IQ}\) and show that it is equivalent to a proper score.

Now we apply the same reasoning to the observations as to the forecasts supposing that the location inside of the neighborhood is forgotten, as it is in the verification of deterministic forecasts with neighborhood (Stein and Stoop 2019). Thus, the local estimator of the CDF for the observations is no longer a function of Heaviside \(H(y - y_i)\) where \(y_i\) is the observation value at the central point but the neighborhood pooled CDF \(G_n\) given by

\[
G_n(y) = \frac{1}{N_n} \sum_{i=1}^{N_n} H[y - y(i)],
\]

where \(y(i)\) \(i = 1, \ldots, N_n\) correspond to the \(N_n\) values observed at the different points of the neighborhood around the central point. The generalization to the energy distance provided by Eq. (5) between forecast and observed CDFs is necessary to include the case where the local estimators of the distribution of the observations are provided by a pooled distribution and not by a point measure. The resulting expression of the \(d_{IQ}\) is called CRPSuno because it uses the proper divergence function associated to the CRPS equal to the unfair integral formulation (“\(u\)”) and “no” in reference to the regional character of the verification as in Ebert (2008). It is given by

\[
\text{CRPSuno}(F, G_n) = d_{IQ}(F_n, G_n). \tag{12}
\]

In using the Eq. (5), where \(X, X'\) refer to the forecasts and \(Y, Y'\) to the observations, we obtain the following:

\[
\text{CRPSuno}(F, G_n) = \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{i=1}^{M} \sum_{j=1}^{M} [x(i, m) - y(j)]
\]

\[
- \frac{1}{2} \left[ \frac{1}{N_n^2} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \sum_{m=1}^{M} \sum_{n=1}^{M} x(i, m) - x(j, n) \right]
\]

\[
+ \frac{1}{N_n} \sum_{i=1}^{N_n} [y(i) - y(j)]. \tag{13}
\]

\(\text{CRPSuno}\) is then computed by the average over the different points and times of the verification sample in the same way as Eq. (3) used for CRPS.

The energy formulation of the calculation of the CRPSuno requires even more calculation resources when the neighborhood is large but its algorithm can be parallelized or optimized to use the redundancy of calculations from one point to another (a FORTRAN 90 routine parallelized through standard open MP directives is available from the authors).

By construction the CRPSuno quantifies the quality of the ensemble to forecast the neighborhood distribution of the observed field \(y\) at the scale of the neighborhood. We should therefore naturally expect to observe a reduction of the double penalty for the forecasts, particularly those of high resolution.

The CRPSuno is based on the comparison of the observed and forecast CDFs which take into account the whole range of values for the outcome variable whereas the SCRPS (Zhao et al. 2021) refers to a binary event. The CRPSuno and the SCRPS are therefore very different in nature. The link between the SCRPS [Eq. (1)] and the CRPS is provided by the use of the energy formulation of the CRPS [Eq. (6)] but applied to the neighborhood frequencies \(I_o\) and \(I_n\) instead of the CDFs. Moreover, the treatment of the observation and forecast neighborhood frequencies is not symmetric for the SCRPS as it is for the CRPSuno. This was possible for the CRPSuno by taking into account the probabilistic aspect of the observations which leads to the addition of a supplementary term related to the observations dispersion in the neighborhood [Eq. (5)].

The deterministic limit CRPSuno_{det} is obtained by taking the case of the ensemble of one member, amounting to

\[
\text{CRPSuno}_{det}(F, G_n) = \frac{1}{N_n^2} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} [x(i) - y(j)]
\]

\[
- \frac{1}{2} \left[ \frac{1}{N_n^2} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \sum_{m=1}^{M} \sum_{n=1}^{M} x(i, m) - x(j, n) \right]
\]

\[
+ \frac{1}{N_n} \sum_{i=1}^{N_n} [y(i) - y(j)]. \tag{14}
\]

We obtain a third term in relation to the CRPSuno_{det} corresponding to the dispersion of observed values in the neighborhood. This score is not a score used classically to evaluate the deterministic forecasts but it allows in a very natural way the introduction of the spatial tolerance in a score based on the AE. The neighborhood frequencies of observed and expected events compared by Roberts and Lean (2008) for binary outcomes are replaced by the CDFs of forecasts and observations pooled in the neighborhood for continuous outcomes.

c. Fair formulations

Zamo and Naveau (2018) have shown that the fair formulation of the CRPS proposed by Ferro (2014) can be obtained in taking nonbiased expressions of the term of dispersion \(\mathbb{E}_{XX'}\). We obtain therefore the fair version of the CRPSuno for the case noted CRPSiso and of its deterministic limit CRPSiso_{det}:
Cases with forecasts perfect but misplaced

The forecast of this field is realized by a set of 16 members which also forecast a zero field across the whole of the forecasting area of $22 \times 22$ points except in one point where the 16 values are also drawn from the same standard normal distribution. The point where the forecast is not zero is confused with the observed field for EXP0, shifted from $N$ points for EXPN for $N$ ranging from 1 to 4 (Fig. 1a).

We evaluate the four neighborhood scores for these five experiments for a neighborhood reduced to one point over the verification area corresponding to the $20 \times 20$ central points. We then consider a neighborhood of $N_n = 3 \times 3$ points for which we keep the same verification domain with the aim of calculating forecasts and observations without additional assumption on the edge points of the verification area.

We start by the verification in relation to neighborhood observations. If the neighborhood is reduced to one point only, we find the same average CRPSuno for all nonzero offsets equal to CRPSu_{0x1}(EXP1) and a smaller average CRPSuno for the case where the offset is zero noted CRPSu_{0x1}(EXP0) (Fig. 1b) since the forecast does not make a location error in this case. We can show, by explicitly counting nonzero cases, that CRPSuno_{0x3} representing the average CRPSuno for a neighborhood of $3 \times 3$ points, can be expressed as the linear combination:

$$CRPSuno_{0x3}(EXPN) = \alpha_N CRPSu_{0x1}(EXP0) + (1 - \alpha_N) CRPSu_{0x1}(EXP1),$$

where $\alpha_N$ is equal to the number of points around which the neighborhood contains at once an observation and a nonzero forecast, divided by $N_n^2 = 81$ for the experiment EXPN. We find $\alpha_0 = 9/81$, $\alpha_1 = 4/81$ and $\alpha_2 = \alpha_3 = \alpha_4 = 1/81$. This linear combination allows us to find the values of the average CRPSuno_{0x3} plotted in Fig. 1b leading to a global decrease in the average score thanks to the taking into account of the perfect forecasts in the neighborhood of $3 \times 3$ points and also to the neighborhood pooling of the observations. We also see that the growth of the error with the shift along the diagonal is slower for a neighborhood of $3 \times 3$ points than for a neighborhood of $1 \times 1$ points due to the spatial tolerance introduced mainly for EXP1. The average CRPSfno is smaller than the average CRPSuno for all EXP0–4 and for all neighborhoods because the CRPSfno is not biased and the CRPSuno is positively biased (Zamo and Naveau 2018). Their difference is important and comes from the correction of the term stemming from the dispersion of observations which is constituted of a small ensemble of $N_n = 9$ members. If we verify these forecasts against the local observations in the case of neighborhoods $3 \times 3$, the average CRPSuso and the average CRPSfso remain much closer as the correction is carried out only on the term for the dispersion of forecasts which is weak because it involves $N_n M$ members corresponding to a set of large size. EXP0 and EXP1 have the same average CRPSfso and average CRPSuso for a neighborhood of $3 \times 3$ points because the offset of 1 point changes nothing in terms of the neighborhood forecasts on the 9 points surrounding the center where the errors are not zero. This value is greater than CRPSu_{0x1}(EXP0) because the neighborhood
forecast is not as good as the local forecast when one compares it to the local observation for EXP0. The neighborhood forecast is on the contrary better for EXP1. For the experiments EXP2 to EXP4, we obtain the same error which is a little higher than for EXP1 as no compensation is produced for a neighborhood of $3 \times 3$ points.

We also observe that most of the score reduction in this example comes from using neighborhood observations rather than from moving from raw forecasts to neighborhood pooled forecasts as shown for instance by the differences for EXP0 to EXP4 between the average CRPSuno and the average CRPSuso for $3 \times 3$ neighborhoods and the average CRPSfso for $3 \times 3$ neighborhoods.

To conclude, these experiments show that the CRPSuso and the CRPSfso quantify the reduction of the error of pooled ensemble forecasts in relation to the case without neighborhood thanks to the spatial tolerance introduced by the mix of forecasts of the different neighboring points. Moreover, a supplementary reduction is quantified by the CRPSuno and CRPSfno when the central observations are also replaced by the neighborhood pooled observations.

b. Case with a uniform perfect forecast

In the idealized UNIF experiment, a uniform numerical setup is obtained by considering a simulation domain of $110 \times 110$ points, where an ensemble forecast of 16 members is compared to a reference: the observation and the 16 forecast values are drawn randomly and independently at each point from the same standard normal distribution. The verification domain corresponds to $100 \times 100$ points in the center of the simulation area so as to eliminate the effects of the lateral boundary conditions in the score variation when the neighborhood size changes. The process is iterated 1000 times. When the neighborhood is reduced to one point, the CRPSn0 (for either CRPSuno or CRPSfno) and the CRPSfso (for either CRPSuso or CRPSfso) are equal by construction. As for the previous experiments, the fair values of average CRPS are smaller than the unfair values of the average CRPS (Fig. 1c). This difference decreases as the neighborhood grows.

For the verification against neighborhood observations, we note a sharp drop of the CRPSuno and the CRPSfso in Fig. 1c because the CDF of the neighborhood pooled forecasts (from $N_nM$ values) compare even better to CDF of neighborhood pooled observations (from $N_n$ values) as the number of neighborhood points involved increases. The unbiased estimator of the neighborhood-based CRPS can be slightly negative due to sampling errors because only the unfair versions of the CRPS are numerically equal to the integral form of the CRPS [Eq. (2)] which is always positive. The bias removal in the energy formulation of the CRPS can lead to very small negative values when the limit value is zero for infinity size of the ensemble (Zamo and Naveau 2018).

For the verification against observations in the center of the neighborhood, we find far smaller variations according to the size of the neighborhood. In fact the mean value for the neighborhood $N_n = n \times n$ of the CRPSuso can be found once more by calculating the average CRPSu without neighborhood for an experiment with an ensemble forecasts which have $16N_n$ members. This result is coherent with the definition of the CRPSuso [Eq. (9)]. This leads to a very small variation of the average CRPSuso with the increase of neighborhood size equivalent to an increase of the ensemble size as in Fig. 1 of Zamo and Naveau (2018). Finally, we note again a smaller variation for the average CRPSfso than for the average CRPSuso as a function of $N_n$, as the CRPSfso is a nonbiased estimator of the unknown value of the CRPS for this experiment and thus less sensitive to the effective size of the ensemble.

In conclusion, we have here the quantification of the impact of enlarging the effective size of the ensemble forecasts by pooling the forecasts at neighboring points either to
improve (or not) the forecast CDF at the center of the neighborhood or to better fit the neighborhood pooled observed CDF.

4. Applications to real cases

a. Description of models and observations

For this study, we use four systems of operational forecasts at Météo-France:

1) The global deterministic hydrostatic model ARPEGE using a stretched calculation grid which goes from 5 km across France to 24 km across New Zealand (Courtier et al. 1991). Its initial conditions are provided by the 4DVAR assimilation cycles across time windows of 6 h (Desroziers et al. 2003). In this study, outputs are oversampled on the regular latitude–longitude grid at 0.025° across western Europe common to all models.

2) The ensemble version of the model ARPEGE, known as PEARP, is constituted of 35 members with a horizontal resolution reduced to 7.5 km across France and 32 km across New Zealand (Descamps et al. 2015). These initial states are obtained by a mix of singular vectors and of perturbed members of an ensemble of fifty assimilations at 40 km. The model error is represented by a random draw for each member among a set of ten coherent and different physics. In this study, outputs are oversampled on the common grid in the same way as ARPEGE outputs.

3) The deterministic nonhydrostatic limited area model AROME (Seity et al. 2011) using a grid of 1.3 km across western Europe. The lateral conditions are provided by the model ARPEGE and the initial conditions come from a cycle of 3DVAR hourly assimilations. In this study, outputs are bilinearly interpolated on the regular latitude–longitude grid at 0.025° across Europe common to all models.

4) The ensemble version of the model AROME is known as PEAROME. It is made up of a set of 16 members having a horizontal resolution of 2.5 km across western Europe. Their lateral conditions are provided by a selection of members of PEARP. Their initial conditions come from an ensemble of assimilations of 25 members at 3.5 km centered around the 3DVAR operational analysis of the deterministic model AROME (Bouttier and Raynaud 2018). The tendencies of certain physical parameterizations are perturbed randomly by a multiplicative factor to represent the error model. In this study, outputs are interpolated on the common grid in the same way as AROME.

Rain observations are provided by the data fusion product ANTILOPE (Laurantin 2008) between radar data of the French radar network and the rain gauges (Fig. 2). The grid of this analysis is of 1-km resolution and the data are averaged on the common grid at 0.025° of forecasts in order to be comparable. The verification domain common for all the four models corresponds to all the points where ANTILOPE is defined (Fig. 2). The comparison between observations and forecasts is realized with the forecasts starting at 0000 UTC for ARPEGE and PEARP as well as the forecasts starting at 0300 UTC for AROME and PEAROME, which use
asynchronous hourly coupling files coming from ARPEGE and PEARP of 0000 UTC. The verification period spreads across a duration of 3 months from October to December 2019. No minimum number of observations present in the neighborhood around the central point is imposed to compute the CRPS at this point. The mask for observed missing rainfall data is also applied to the forecasts so as to permanently keep the same number of points for the forecasts and observations in every neighborhood. This means that the number of points in the neighborhood varies when masked observations are present. The computation of the neighborhood scores is performed only when all the points of the neighborhood are inside the common verification domain.

So as to see the impact of the distribution law of several outcome variables, we also consider the 2-m temperature. The observations are also provided by a data fusion product named APROFUS (Marcel and Laurantin 2019) which mixes the 2-m temperature observations of the observational network operational across France with the AROME analysis on a kilometer grid. The value of the observation is obtained in the center of each forecast grid by the bilinear interpolation of the four APROFUS points which surround it. The verification domain common to all the four models for 2-m temperature corresponds to the APROFUS domain plotted in Fig. 2 except at the border, while the neighborhood scores are computed only if all the points of the neighborhood are inside the APROFUS domain.

So as to evaluate the statistical significance of observed differences between two forecast systems, a block-bootstrap technique (Hamill 1999) is applied to time series choosing randomly 1000 times with delivery of the 3-day blocks among the 92 days of the verification period. The level of the test is fixed to 5% for all the six possible comparisons of the four systems.

**b. Quarterly verification of cumulative rainfall over 3 h**

We will start by analyzing the results of these comparisons with the CRPS when the neighborhood is reduced to one grid point. The CRPS is equal to the fair version of the CRPS (Ferro 2014) and the deterministic version corresponds thus to an AE. The deterministic models AROME and ARPEGE have similar average scores (Fig. 3a). The analysis of AROME seems more accurate than that of ARPEGE as an initial condition for short term rain forecasts. This can probably be explained by the assimilation of radar data which is only present for AROME. However, from 18 h of forecasts, the mean AE are equivalent and we even observe across the last lead times an advantage for ARPEGE. The average CRPS of ensemble forecasts are significantly far better than the mean AE of deterministic forecasts showing a real benefit in the evaluation of CDF observed by ensemble forecasts in relation to a deterministic forecast even at higher resolution. The growth rate of the forecast error measured by the average CRPS is between 2 and 3 times greater for the deterministic forecasts than for the ensemble forecasts. Finally, PEARP and PEAROME provide close values of the CRPS.

If we introduce a neighborhoods of size 0.125° by including 5 × 5 points in the comparison between these four forecasts, all the models have better average CRPS (Fig. 3b). This indicates that the double penalty is reduced when the neighborhood pooled CDFs replace their local estimations allowing compensations between members and localizations inside the
neighborhood. This is clearly more active for deterministic forecasts than for ensemble forecasts and leads to a reduction of the gap between their average CRPSfnocurves. The neighborhood methodology is therefore validated again for the deterministic forecasts for continuous outcomes in the same way as with FSS (Mittermaier et al. 2013) or neighborhood-based table of contingency (Stein and Stoop 2019) for categorical outcomes. It can be seen that the average CRPSfnos of AROME is now significantly better than the CRPSfnos of ARPEGE for all lead times indicating that the double penalty correction by neighborhood treatments is more active for a high resolution model as expected. The neighborhood treatment also works well for ensemble forecasts and helps to quantify how neighborhood probabilities are better forecast by ensemble forecasts than local probabilities: the average CRPSfnosdet of ensemble forecasts is reduced by 0.07 mm for all lead times when the neighborhood includes 5 × 5 points (Fig. 3b) instead of 1 × 1 point (Fig. 3a). Moreover, the average CRPSfnosdet of the deterministic forecasts grow faster with the lead time than the average CRPSfnos of ensemble forecasts in the same proportion as with a neighborhood reduced to one point.

When we enlarge the neighborhood to 0.275° (11 × 11 points), all the models have better scores (Fig. 3c). We note again nonetheless the same conclusions as for the previous neighborhood: the AROME analysis allows an improvement of the AROME forecasts on short lead times by getting closer to the quality of the ensemble forecasts but these last ones are then significantly better due to the slower speed of the error growth. According to this indicator PEAROME and PEARP are of equivalent quality for this scale even if the small but significant difference is in favor of PEAROME. AROME performs significantly better than ARPEGE for all lead times.

So as to have a more synthetic view of the dependence of scores with the size of the neighborhood, we select two lead times at 6 and 42 h so as to have a lead time greatly influenced by the initial conditions and a lead time farther away (Fig. 4). We can measure the performance of the AROME analysis across all the scales when we compare ARPEGE and AROME forecasts on a short lead time (Fig. 4a). This is no more the case at longer lead times (Fig. 4b) and the forecasts become very close even if the use of neighborhoods greater than one point reduces significantly more the average CRPSfnos for AROME than for ARPEGE. The predictive content for AROME also outperforms the PEARP content at 6 h for scales greater than 0.0525° or 21 × 21 points according to this score (Fig. 4a). However, we see that on a long lead time, the ensemble forecasts provide an undeniable benefit converging between themselves for all scales because the growth of the error here is far slower than for the deterministic forecasts (Fig. 4b).

We plot on the Fig. 5 the time series of the daily average CRPSfnos and the averages of daily rain observed across the verification area. We note the very strong time correlation between these two quantities which shows that the rainiest cases are also the least well predicted. In contrast the ranking of the
four forecasts between themselves remains very stable. It depends little on the observed rain and it is coherent with the mean rank shown at lead time 18 h on Fig. 4b. It is a good property of this score which shows its capacity to compare deterministic and probabilistic forecasts for a given scale whereas the forecasts use different resolutions as well as the possibility to exploit this score on a daily basis.

Any improvement from the neighborhood pooling postprocessing can be quantified by a comparison of the average CRPS to the average CRPS of the raw ensemble forecast. We plot for the average CRPS of the Fig. 6 the equivalent for the average CRPS of the Fig. 4. The comparison of these two figures shows that the dependence on the neighborhood is very different with a quasi-saturation for the average CRPS for the large neighborhoods whereas the average CRPS decreases regularly with the size of the neighborhood. The difference between these two behaviors comes by design of the taking into account of the dispersion term of

![Figure 5](image1.png)

**Fig. 5.** Daily temporal series of the average CRPS (in mm) for the rain accumulated during 3 h for PEARP (light gray full line), PEAROME (black full line), ARPEGE (light gray dashed line), and AROME (dark gray dashed line). The average CRPS are computed with neighborhood sizes equal to 0.125° at 1800 UTC on day D. The daily temporal series of the rain accumulated during 3 h observed by ANTILOPE and averaged over the whole domain is also superimposed with a dotted line.

![Figure 6](image2.png)

**Fig. 6.** As in Fig. 4, but for the average CRPS for the rain accumulated during 3 h.
observations in the neighborhood which reduces the CRPS [Eqs. (9) and (13)] and allows the average CRPSfno to be smaller than the average CRPSfso for large neighborhoods. This corresponds to a profound process in the verification which is to take into account the same spatial scale for the observations and the forecasts. This approach is in the same spirit as the taking into account of the errors of observations (Ferro 2017) or indeed the errors of representativity (Ben Bouallegue et al. 2020). Future studies will analyze how these approaches can merge in this formalism. It is not easy to interpret the quasi-saturation of the average CRPSfso for the large neighborhoods because we could imagine that the points that are far away must have a negative contribution to the CDF of the ensemble forecasts because they are decorrelated from the central point where the observation is taken into account while the contribution remains almost zero from these points. One possible interpretation of the quasi-saturation of the average CRPSfso is that the observed and predicted precipitation fields are poorly correlated for small neighborhoods due to the large variance. As a result, the correlation errors are already saturated at small scales, and any decorrelations in the large scale precipitation that result from large forecast neighborhoods do not affect the average CRPSfso. Reductions in the forecast variance reduce the average CRPSfso score for small to medium sized neighborhoods, but the variance becomes vanishingly small for large neighborhoods, limiting its effect. It can be noted that deterministic and ensemble forecasts give very close values of the average CRPSfso for the large neighborhoods showing the relevance of replacing the local deterministic information by a smoother flow of information corresponding to neighborhoodized frequencies built with neighboring points.

The plots of the other two unfair average scores CRPSuno and CRPSuso (not shown) differ from their average equivalents CRPSfno and CRPSfso by only 1% at most. It shows the little impact of the correction of the normalization of dispersion terms in real cases. These variations are therefore completely negligible compared to those generated either by the passage from “so” to “no,” either by the variation of the neighborhood or by the variation of the lead time, at least for the ensemble sizes considered.

c. Quarterly verification of 2-m temperatures

The aim of this section is to show the impact of the distribution of the outcome variable or its relation to orography, on the variation of scores with the neighborhood size. For this, we plot the variations of the CRPSfno and the CRPSfso according to the size of the neighborhood for the 2-m temperature (Figs. 7 and 8), which are to be compared to their equivalents for the accumulated rain across 3 h (Figs. 4 and 6). For the CRPSfno, we note that the impact of the resolution is clearer for the forecast of 2-m temperature than for the rain as even with large neighborhoods we observe neither the convergence of scores of the ensemble forecasts PEAROME and PEARP on a long lead time nor the convergence of the AROME and ARPEGE deterministic forecasts. Instead, we observe a grouping by system resolution, with the high-resolution PEAROME and AROME systems significantly outperforming the low-resolution PEARP and ARPEGE systems at the 5% significance level. Surely AROME and PEAROME have an advantage over ARPEGE and PEARP since APROFUS uses AROME analyses as background to be combined with station data. Moreover the ensemble systems perform substantially
better than their respective deterministic counterparts when both the lead time is long and the neighborhood size is small.

The dependence on the size of the neighborhood is very different for the CRPSfso of the 2-m temperature since the taking into account of the neighboring points improves the local forecast only for the neighborhoods of small size and frankly degrades it for larger neighborhoods (Fig. 8). This can be explained by the existing decorrelation for the 2-m temperature in a given point and its neighbors, which can be at very different heights and thus little representative of the uncertainty other than that which is local. It illustrates that neighborhood averaging, if not done carefully, can add biased members to the ensemble. Such a behavior also exists in the case of the CRPSfno as the possibility to recover among the neighboring points the forecasts closer to 2-m temperature in relation to an observation at a given location is reduced as soon as the orography varies significantly. However in this case, we treat coherently the forecasts and the observations, which allows us to keep an informative score depending on the scale analyzed.

5. Discussion of the results

The CRPSfso could be preferred for an end user who is mainly interested by the forecast at a given location. In this case, the CRPSfso helps to quantify the improvement brought by the forecasts at neighboring points to the forecast at the central point. It has been shown that this improvement can be negative in the case where the observations strongly depend on the location as for the temperature case and that the CRPSfso warns you about a bad use of neighborhood as a posttreatment in this configuration. Moreover, the CRPSfso could reward forecasts aggregated and smoothed coming from neighboring points more than forecasts coming from the statistical distributions of the observations taken at neighboring points and therefore leading to counterintuitive results. Indeed, the propriety of the CRPS only says that the best forecast is the one extracted from the statistical distribution of the observations at the same central point. It does not say anything about the optimality of a forecast built from different statistical distributions of the observations taken at different points of the neighborhood.

The CRPSfso has the disadvantage to compare ensemble forecasts and observations at two different scales and following Mittermaier and Csima (2017), we should try to limit the size of the neighborhood to a maximum value, under which the assumption of equiprobability of the observations in the neighborhood is not broken in order to use the neighborhood methodology to reduce the double penalty. This is clearly not the case for temperature over France where high mountains are present. One supplementary question comes from the overlapping of neighborhoods when we try to use the CRPSfso as a metrics to minimize in order to statistically postprocess the forecasts. The forecasts at one point will belong to multiple neighborhoods and could be submitted to contradictory constraints from observations coming from different statistical distributions. For this reason, it is prudent to limit the use of the CRPSfso to diagnostic posttreatments with reasonable sizes.

The CRPSfno compares the ensemble forecasts aggregated over the neighborhood to observations aggregated in the same way. The propriety of the $d_{100}$ guarantees that the best ensemble forecast at the scale of the neighborhood will be the

![Fig. 8](image-url)
one extracted from the statistical distribution of the observations aggregated over the neighborhood. Even if the statistical distribution of the observations varies inside the neighborhood, the CRPS_{no} is able to indicate how the forecasts neighborhood distribution fit the observations neighborhood distribution. This property pleads for the CRPS_{no} which allows you to objectively compare at a given scale a forecast with observations and even forecasts of different resolutions with the same reference. One needs to be careful when comparing the “no” scores across neighborhood sizes. Since the type of observations changes with the neighborhood size, the true best expected score also changes. Thus, a smaller “no” score for a neighborhood size S2 than for a neighborhood size S1 can be due either to a better pooled ensemble forecast at the size S2 than at the scale S1 or to a better predictability of the pooled observations at the scale S2 than at the scale S1.

The choice of a meaningful neighborhood size for the CRPS_{no} is not easy because it clearly depends on the outcome variable. There is no systematic asymptotic value for large neighborhoods for the CRPS_{no} as reported by Roberts and Lean (2008) for the spatial verification of deterministic forecasts of binary events. Nevertheless, the value of the neighborhood size can be fixed by an external constraint like in Amodei et al. (2015) to be representative of an administrative department or in Taillardat and Mestre (2020) to realize a balance between the representativity of rain and the numerical cost.

6. Conclusions

We perform an implicit postprocessing step by linearly pooling the forecasts of different members of different points of the neighborhood to make a superserene realizing a forecast on the neighborhood scale. The CRPS is then used to compare the pooled forecast CDF with that provided by the only observation in the center of the neighborhood. This CRPS expression for pooled ensemble forecasts is called CRPS_{so} and it quantifies therefore the relevance of using the points of neighboring forecasts to improve the local probabilistic forecast. We also compare this pooled forecast CDF to that corresponding to all the observations collected in the neighborhood by using the divergence function \( d_{40} \) derived from the CRPS. This second CRPS expression for pooled ensemble forecasts is called CRPS_{no}. This score quantifies the quality of the forecast on the scale of the neighborhood considered. Fair versions of these two neighborhood scores have also been constructed so as to reduce the bias for the ensembles of small size when forecasts and observations are issued from the same statistic law in following the recommendations of Ferro (2014) and Zamo and Naveau (2018). The extensions of these neighborhood scores in the deterministic case allow one to generalize the AE coherently taking into account the neighborhoods in the calculation of these scores. We can thus directly compare the generalized AE of deterministic forecasts with these CRPS expressions of pooled ensemble forecasts.

A series of idealized cases has allowed us to show the sensitivity of these scores to spatial error of placement of structures and to the relevance of the identical treatment of observations and forecasts to improve their comparisons. These four neighborhood scores have also been used to compare the forecasts of rain accumulated during 3 h and of 2-m temperature for four systems of deterministic forecasts and probabilistic forecasts operational at Météo-France using different resolutions. We were able to show the superiority of the ensemble forecasts over deterministic forecasts for long lead times due to the slower growths of their forecast errors. The CRPS_{so} helps us to show that the use of neighboring points to improve the forecast of the local probability in the center of the neighborhood is very dependent on the forecast outcome variable with a very rapid deterioration in the case of 2-m temperature and in contrast a rapid saturation for rain for large neighborhoods. Because the CRPS_{no} treats coherently the forecasts and the observations in the neighborhood aggregation, it should be preferred to the CRPS_{so} for the comparison of models at the scale of the neighborhood. The CRPS_{no} variation as a function of the neighborhood size will help to quantify the influence of the double penalty on the relative quality of two ensemble forecasts pooled at the neighborhood scale.

It is very important to evaluate precisely the qualities and the faults of a forecast system not only for the whole range of values but also for selected thresholds. For this, work is in progress to introduce the neighborhood pooling as a first step before the computation of the Brier Score in the same spirit as for the CRPS expressions. In addition, the calculation of these CRPS expressions for pooled ensemble forecasts is going to be even more optimized so as to reduce the high computational cost for the large neighborhoods.

Acknowledgments. We are grateful to Maxime Taillardat and Michaël Zamo for discussions on the divergence functions, to Naomi Riviere for reviewing this manuscript, and to the three anonymous reviewers for their valuable comments.

REFERENCES


