

On the verification of the crossing-point forecast

By ZIED BEN BOUALLÈGUE*, *The European Centre for Medium-Range Weather Forecasts (ECMWF), Reading, UK*

(Manuscript Received 13 October 2020; in final form 1 April 2021)

ABSTRACT

The crossing-point forecast is defined by the intersection between a forecast (conditional) and a climate (unconditional) cumulative probability distribution function. It is interpreted as the probabilistic worst-case scenario with respect to climatology. This article discusses a scoring function consistent for the crossing-point forecast where both forecasts and verifying observations are expressed in terms of a climatological probability level. Scores defined in ‘probability space’ are commonly used for the verification of deterministic forecasts and this concept is here generalised to ensemble forecast verification. Practical challenges for its application as well as the sensitivity of the score to ensemble size (number of ensemble members) and to climatology definition (number of used climate quantiles) are illustrated and discussed.

Keywords: scoring function, probabilistic forecast interpretation, ensemble verification, diagonal score, crossing-point

1. Introduction

Consider a probabilistic weather forecast F issued at a location with climatology G for that time of the year, both expressed in the form of a cumulative probability distribution function (cdf). The forecast F is the conditional distribution based on the information at hand when the forecast is issued, while G is the corresponding *unconditional* distribution function of the same random variable of interest. We set our focus on the point of intersection between forecast and climatology cdfs, F and G respectively. The projection of the forecast–climate intersection onto the probability level axis is called the *crossing-point forecast*. The corresponding crossing-point observation is the observed event expressed in terms of its climatological frequency rather than its absolute value. This transformation, the projection onto the probability level axis, is referred to as a projection in ‘probability space’.

The assessment of the crossing-point forecast requires the design of an error function. The score proposed in this article is not (strictly speaking) a new score, but is directly derived from the *diagonal score*, a scoring rule recently introduced by Ben Bouallègue et al. (2018). This manuscript sheds new light on the interpretation of this score and clarifies the link with a score routinely used in the meteorological community, namely the stable and

equitable error in probability space (SEEPS; Rodwell et al., 2010). SEEPS serves as a headline score for assessing and communicating trends in precipitation forecast performance at the European Centre for Medium-Range Weather Forecasts (ECMWF).

The concept of ‘score in probability space’ was first developed in the context of deterministic forecast verification as an attempt to overcome the pitfalls of traditional scores that generally discourage forecast of extreme values, in particular for skewed-distributed variables such as precipitation (Potts et al., 1996; Ward and Folland, 1991). Rather than comparing forecast and observation in ‘measurement space’, the comparison takes place after projection in ‘probability space’. This manuscript intends to show how this concept can be formulated in a probabilistic forecasting context and applied to the verification of ensemble forecasts.

The manuscript is organised as follows: Section 2 presents the score definition, its properties and its relationship with other existing and better established scores. Section 3 presents applications in terms of crossing-point forecasts, an analysis of the score sensitivity in the context of ensemble forecast verification, as well as the concomitant challenges for the score computation. Following the Conclusion in Section 4, mathematical derivations and scoring algorithm are detailed in the Appendices.

*Corresponding author. e-mail: zied.benbouallegue@ecmwf.int

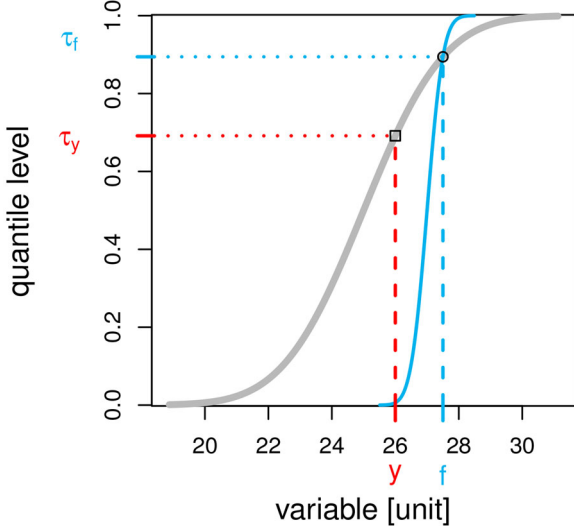


Fig. 1. The necessary ingredients: a climatology cumulative probability distribution function (grey), a probabilistic forecast (blue) and a verification (red). The projection of the circle and square onto the probability space, τ_f and τ_y , are called the forecast and verification crossing-points, respectively.

2. The error function

2.1. Definition

Figure 1 sets the scene. We consider the situation where we have access to the following pieces of information: a climatology known to everyone (or unconditional cdf, denoted G), a probabilistic forecast issued by a forecaster (or conditional cdf, denoted F), and the outcome of a random process (or verification, denoted y).

We consider the following single intersection (SI) condition: *Given F and G two cumulative distribution functions, F and G satisfy the single intersection condition if there exists one and only one f such that:*

$$x \geq f \Rightarrow F(x) \geq G(x) \quad (1)$$

and

$$x < f \Rightarrow F(x) < G(x). \quad (2)$$

In this condition, the intersection point between the forecast F and the climatology G is denoted (f, τ_f) with $\tau_f := G(f)$, while the intersection between verification y and climatology is denoted (y, τ_y) with $\tau_y := G(y)$. We refer to τ_f and τ_y as crossing-point forecast and crossing-point verification, respectively. The crossing-point verification τ_y is the climate quantile level corresponding to the observation y , while the crossing-point forecast is introduced as a new type of probabilistic forecast. Illustrations of crossing-point forecasts and corresponding verification based on synthetic and real data are provided in Fig. 1 and Section 3, respectively.

In the SI condition, τ_f and τ_y are the unique projections of the forecast F and verification y , respectively, in probability space. In this context, the question that arises is how to define an error function that can be applied to a forecast τ_f and a verification τ_y in order to assess crossing-point forecasts appropriately.

Our proposition is the following. Given a forecast τ , an observation y and the distribution G , the scoring function

$$S_G(\tau, y) = \begin{cases} G(y)^2 - \tau^2 & \text{if } G(y) \geq \tau \\ (1 - G(y))^2 - (1 - \tau)^2 & \text{if } G(y) < \tau \end{cases} \quad (3)$$

is a consistent scoring function for the crossing-point functional $T_G : F \rightarrow \tau_f$.

For convenience, we note the corresponding score simply S with $S(\tau_f, \tau_y)$ the result of the comparison between a forecast τ_f and a verification τ_y following Eq. (3). By contrast, we also show why more simplistic error functions, such as for example a naive score defined as the squared difference $(\tau_f - \tau_y)^2$, are not appropriate for the verification of probabilistic forecasts.

2.2. Illustrations

But first, let's get familiar with the error function defined in Eq. (3). Figure 2 illustrates how $S(\tau_f, \tau_y)$ evolves as a function of the relationship between crossing-point forecast τ_f and crossing-point verification τ_y . In Fig. 2a, τ_y is fixed and τ_f varies in the interval $[0, 1]$. This plot allows us to better visualize how a forecast is penalized given a verification. For example, when $\tau_y = 0.5$, the penalty of the score S increases rapidly with departures of τ_f from the verification (solid line). In other cases, when $\tau_y < 0.5$ for example, we see how under-forecasting ($\tau_f < \tau_y$) is less penalized than over-forecasting ($\tau_f > \tau_y$) in this type of situation (dotted and dashed lines). By symmetry, the opposite is true when $\tau_y > 0.5$ (not shown).

In Fig. 2b, we see how S varies as a function of τ_y in $[0, 1]$ for given τ . For the interpretation of this plot, we can recall that, by definition, the verification τ_y is uniformly distributed on $[0, 1]$. One interesting aspect is that the integral of the error function S over all $\tau_y \in [0, 1]$ is independent of the fixed forecast. When the same probabilistic forecast is issued every time (τ_f is a constant), the mean score S over all possible cases does not depend on the forecast. This score propriety is called *equitability* (Gandin and Murphy, 1992). So S is said to be *equitable* because the expected value of the score S is the same for all non-informative (constant) forecasts (see Appendix A for a formal illustration).

A second set of illustrations is provided in Fig. 3. This time, the aim is to illustrate the score sensitivity to predefined typical forecast discrepancies in a controlled environment. For this purpose, the simple toy-model proposed

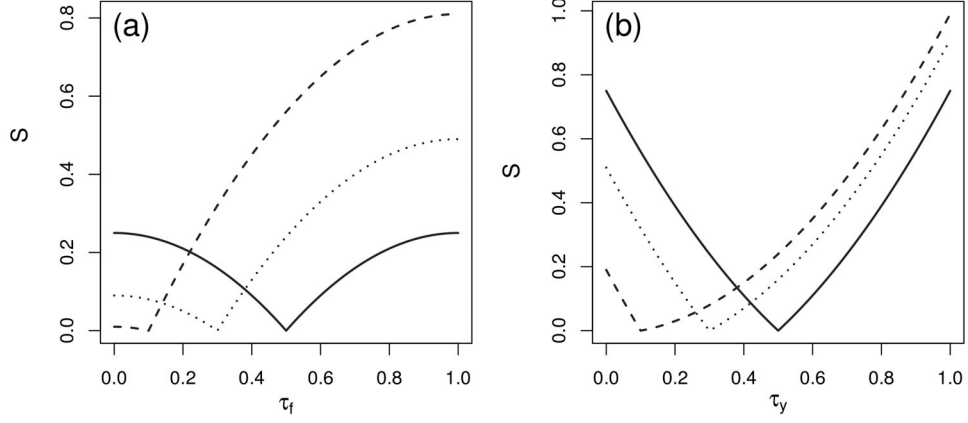


Fig. 2. Getting familiar with the error function. (a) S as a function of τ_f for three different values of τ_y : 0.1 (dashed line), 0.3 (dotted line) and 0.5 (solid line). (b) S as a function of τ_y for three different values of τ_f : 0.1 (dashed line), 0.3 (dotted line) and 0.5 (solid line).

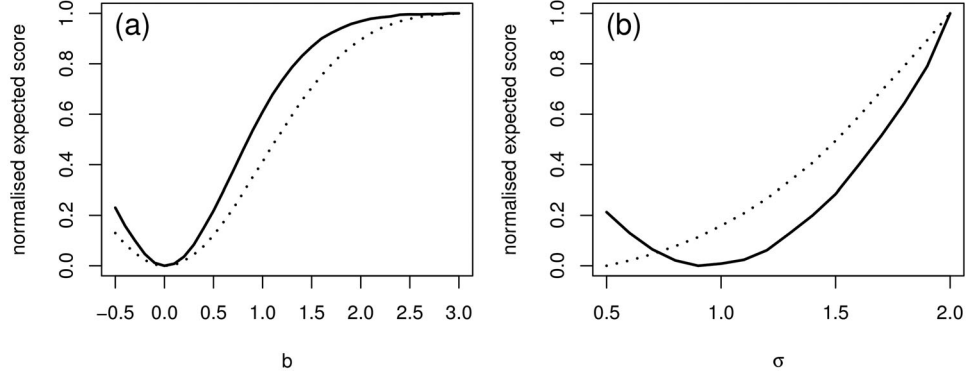


Fig. 3. Comparing two scores in ‘probability space’ using a toy-model while varying the forecast bias (a) and the multiplicative spread factor σ (b). Normalised expected value of the score S (solid line) is compared with normalised expected value of a naive score in probability space computed as $(\tau_f - \tau_y)^2$ (dotted line).

by Lerch et al. (2017) is used. Observations and forecasts are drawn from the same normal distribution:

$$\mathcal{N}(\mu, \alpha^2) \quad \text{with} \quad \mu \sim \mathcal{N}(0, 1 - \alpha^2), \quad \alpha \in (0, 1) \quad (4)$$

where large (small) values of the parameter α indicate low (high) predictability. The climatology is the normal distribution $\mathcal{N}(0, 1)$. For now, and without loss of generality, we set $\alpha = 0.5$. So far, the observation is statistically indistinguishable from a draw of the forecast distribution.

Two experiments are performed in order to make the probabilistic forecast deviate from perfect calibration. First, synthetic data sets are generated adding a bias b , varying in $[-0.5, 3]$, to the mean forecast. Second, data are generated by controlling the level of forecast variance with a multiplicative factor σ . The spread multiplicative factor σ is varied in $[0.5, 2]$. So, the forecast distribution follows a distribution $\mathcal{N}(b, \sigma^2)$, where perfect calibration corresponds to $b = 0$ and $\sigma = 1$ in each experiment. Normalised scores as a function of b and σ are plotted in Fig. 3a,b, respectively.

In Fig. 3, we compare the forecast performance, as b and σ vary, considering two scores: S and the naive score (squared difference) in probability space. For both scores, the minimum is reached when $b = 0$. A saturation for large biases is also visible with S reaching a plateau when $b > 2$ (the bias exceeds two times the climate standard deviation). However, when varying the forecast variance, S reaches its minimum when $\sigma = 1$, that is when the forecast is perfectly calibrated, while the naive score has its minimum for the lowest tested value of σ . So, on one side, this plot clearly illustrated that the naive score favours forecast that are not well-calibrated and so can be deemed as inappropriate for forecast comparison. On the other side, this plot also hints that S proceeds from a proper scoring rule. Further discussion on S properties follows.

2.3. Interpretation

The choice of the error function in Eq. (3) is not fortuitous but results from a derivation of the diagonal score, a

Table 1. Error scoring matrix used both for the derivation of the diagonal score and SEEPS. Forecast and observation refer to an event with climatological frequency of occurrence p_0 .

		Observed	
		Yes	No
Forecast	Yes	0	p_0
	No	$1 - p_0$	0

proper score introduced by Ben Bouallègue et al. (2018). We show in Appendix B that the diagonal score expressed in terms of τ_f and τ_y is equivalent, up to a factor 2, to the score S when the crossing-point forecast exists and is unique. In the original study cited above, the diagonal score is interpreted as a score ‘tailored to vulnerable users, [...] those more exposed to stress as the weather event severity increases’. More formally, it is defined as the integral of the *diagonal elementary score* over all quantile levels.

As a new result, the diagonal elementary score is established as a proper score for an *interval probabilistic forecast* (Mitchell and Ferro, 2017). An interval probabilistic forecast is for example ‘there is [10–20]% chance of rain tomorrow’. In Fig. 1, it is clear why this type of probabilistic forecasts is relevant here. For any event (defined as the variable of interest exceeding a threshold), a probability forecast is reduced to an interval probabilistic forecast. Focusing on the crossing-point, the only information retained from a probability forecast, say $p_f = 1 - F(t)$, where t is the given threshold defining an exceedance event, is whether this probability p_f is greater or lower than the climatological probability of occurrence denoted p_0 (with $p_0 = 1 - G(t)$). So, for any binary event, we focus on a forecast which takes the form of a probability interval: $[0, p_0]$ or $(p_0, 1]$. In Appendix C, we show that a proper score for such an interval probabilistic forecast is the diagonal elementary score characterised by the error matrix in Table 1.

The error matrix in Table 1 indicates the asymmetrical penalties associated with the diagonal elementary score. This error matrix is key to understanding the relationship between the diagonal score and decision-making based on a standard cost/loss model (a detailed discussion on that point can be found in the work by Ben Bouallègue et al., 2018). This table also explicitly shows the relationship between S and SEEPS. Table 1 is equivalent, up to a constant factor $p_0(1 - p_0)$, to Table X in the work by Rodwell et al. (2010) which shows the ‘two-category equitable error matrix for a score that SEEPS can be built from’. While SEEPS focuses on three categories (dry weather, light rain and heavy rain), the score S is built on a finer (possibly exhaustive) description of the climate

distribution. Sensitivity of the score S to the climatology definition in terms of quantiles is discussed in Section 3.3.

3. Applications

3.1. Crossing-point forecasts

A crossing-point forecast is derived from the comparison of a conditional with an unconditional probability distribution. The two cdfs are summarised into a single number, one characteristic of a probabilistic forecast. This number provides information about the forecast level of risk with respect to the climatological level of risk. There is no focus on one particular event (as for example the risk of having temperature exceeding 30°C) or on a particular climate quantile (as for example the 95% percentile), but rather a scanning of all possible events/quantile levels. In the SI condition, the crossing-point corresponds to the pivotal point where the probability forecast for an exceedance event switches from higher to lower than climatological frequency. So, in plain words, the crossing-point forecast is associated with the *worst-case scenario* which is more likely based on the information at hand than without, the largest threshold so that the corresponding exceedance event gets assigned an above-climatological probability based on the current probabilistic forecast. By convention, we express the crossing-point forecast in terms of a quantile level (a number between 0 and 1) but it could also be communicated in terms of a return-period (Prates and Buizza, 2011) or a quantile value (Hawkins and Kochar, 1991).

In Appendix D, we argue that the scoring function S_G is consistent for the crossing-point forecast, and the diagonal score is the corresponding proper score. The idea of *consistency* between a score and a forecast directive (or functional) relates to the concept of *elicibility* (Gneiting, 2011). A statistical functional is called elicitable if there is a ‘scoring function or loss function such that the correct forecast of the functional is the unique minimizer of the expected score’ (Fissler et al., 2019). For example, the distributional mean is elicitable with the root mean squared error as a consistent loss function. These mathematical tools and related concepts help drawing robust conclusions from the comparison of competing forecasts in a probabilistic framework.

We illustrate the concept of crossing-point forecast (and its consistent assessment) with an example based on the operational ensemble prediction system (ENS) run at ECMWF. More specifically, we analysed a 2m temperature forecast, in the form of model grid-box averages, considering instantaneous values at 12UTC. The spatial grid resolution of the ensemble forecast is approximately 18km, but the forecast is here interpolated on a

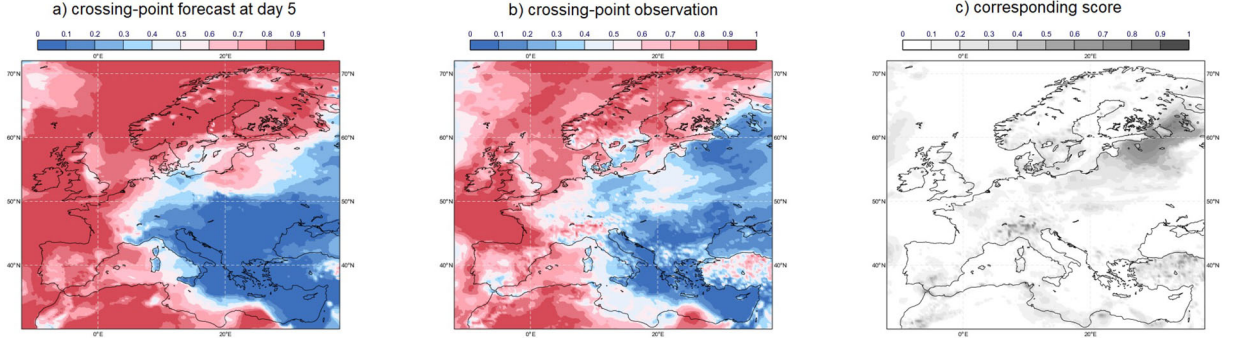


Fig. 4. (a) Two metre temperature crossing-point forecast derived from ENS at day 5, (b) crossing-point observation on 1 June 2020 and (c) corresponding score S .

$0.25^\circ \times 0.25^\circ$ grid. The verification corresponds to the analysis on the validity date. The climatology is a model-climatology derived from reforecasts. Section 3.2 details how crossing-point forecasts are generated.

Figure 4 compares qualitatively (by visual inspection) and quantitatively (by computing S) the forecast and verification crossing-points of 2m temperature valid on 1 June 2020. The focus is on the European-North African domain. In Fig. 4a, the western and northern parts of the domain appears at ‘higher risk than normal’ for large positive anomalies (crossing-point close or equal to 1), while Southern and Central Europe is dominated by a signal of ‘higher risk than normal’ for large cold anomalies (crossing-point close or equal to 0). By comparison with Fig. 4b, we appreciate the overall agreement in terms of spatial structures between crossing-point forecast and verification, but we also note that the verification map displays values more equally distributed over the interval $[0,1]$ than the forecast map. Because S is a scoring function, it can be computed for each pair of forecast and verification (each model grid-point). As shown in Fig. 4c, large-scale poor forecast performance affects only North-Eastern Europe in this example.

3.2. Computation

When τ_f and τ_y are known, the computation of the score S is straightforward. However, this is generally not the case, and, to the best of the author’s knowledge, there is no closed-form for the computation of the intersection point between two cdfs in the case of well-defined distributions such as normal distributions. In this context, two different approaches can be followed: (1) a pragmatic approach to find the crossing-point forecast τ_f , or (2) the direct computation of the diagonal score in its original formulation. Both approaches are discussed below.

The first option is to be pragmatic and to estimate τ_f and compute $S(\tau_f, \tau_y)$ based on Eq. (3). Consider for an example an ensemble forecast with members $e_1, \dots, e_M$

and a climatology defined by unique quantiles $q_1, \dots, q_{n_q}$ at increasing levels $\tau_1, \dots, \tau_{n_q} \in (0, 1)$. The forecast probability of exceedance $p_i, i = 1, \dots, n_q$, is derived by counting the number of ensemble members exceeding the respective quantile q_i , that is, $p_i = \frac{1}{M} \sum_{k=1}^M \mathbb{I}[e_k > q_i]$. Then the crossing-point forecast is found by comparing the exceedance probabilities p_i to the climate probability levels $1 - \tau_i$. Let $j \in 1, \dots, n_q + 1$ be the smallest index i such that $p_i \leq 1 - \tau_i$, with $j = n_q + 1$ if no such index exists. The crossing-point forecast τ_f is set equal to $\frac{1}{2}(\tau_j + \tau_{j-1})$, where $\tau_0 = 0$ and $\tau_{n_q+1} = 1$ are used at the boundaries.

This simple approach is followed to produce the illustrative example in Fig. 4. In our examples, $M = 50$ and τ_i takes value 1%, 2%, ..., 98% and 99%. This approach can be refined by interpolating around the intersection point as illustrated in Fig. 5a. In addition, extrapolation could be performed using extreme value theory for a finer assessment of crossing-points close to the distribution tails. The application of this later step goes beyond the scope of this study.

The second option consists in directly computing the diagonal score which does not require the estimation of τ_f . In order to facilitate the application of this approach for the verification of ensemble forecasts, an algorithm is provided in Appendix E. Besides the ensemble forecast and a verifying observation, the diagonal score computation requires as input a climatology defined by a set of quantiles. The score is computed as the mean diagonal elementary score over all unique climate quantile levels. Climatological distributions of weather variables such as precipitation are censored distributions and the first two loops of the algorithm are dedicated to identifying the set of unique quantile values within the variable bounds. In Section 3.3, we discuss the sensitivity of the score with respect to the ensemble size as well as the climatology definition.

The relationship between S and the diagonal score holds only in the SI condition. However, in practice, multiple intersection points can coexist for a single

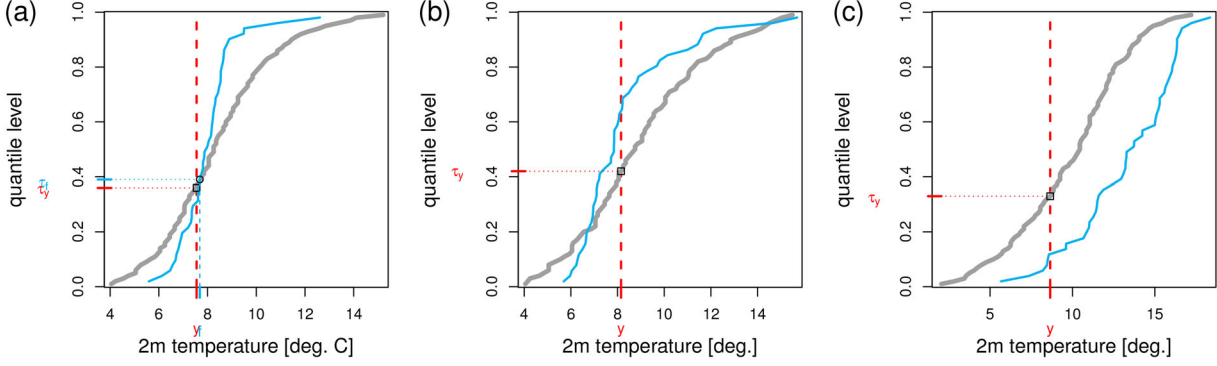


Fig. 5. Same as Figure 1 but based on a real data set: 2m temperature ENS forecasts at day 5 valid at three different stations illustrating: (a) the single intersection condition, (b) a case of multiple (2) intersections and (c) a case with zero intersections. The climatology is site-specific, based on a 30-year observation records covering the period 1980–2009.

forecast–climatology pair as illustrated in Fig. 5b. How often this situation is encountered in real applications is examined with the help of two distinct datasets, one dealing with temperature at 2 m above the ground and one with daily precipitation. Over Europe, focusing on June 2018, pairs of ensemble forecast climate distributions are analysed at approximately 1500 synoptical stations: for each pair, the number of intersection between the forecast and climate distributions is counted. The distribution of the number of intersections per pair is displayed in Fig. 6.

The prevalence of the SI condition in both the temperature and precipitation datasets is illustrated in Fig. 6a,b, respectively. Multiple intersections represent 4% (5%), 11% (13%) and 22% (28%) of all cases at day 2, 5 and 10, respectively, for 2m temperature (daily precipitation). Cases with zero intersections (0 category) are the limit cases of the SI condition: the crossing-point is defined and can in principle take value in $[0, \tau_1]$ or $(\tau_{n_q}, 1]$. An example of a case with zero intersections is provided in Fig. 5c. Based on the results in Fig. 6, we infer that the SI condition is more often associated with forecasts at shorter lead time, that is forecast with a sharper probability distribution. At longer lead time, multiple intersections are more frequent. The number of cases with zero intersections is also rising with the forecast horizon, leading to a larger number of crossing-points taking value 0 or 1. For longer lead time, the forecast can become similar to climatology and a single crossing-point is difficult to identify in that case. In terms of score, S converges to the score value for non-informative forecasts, that is random or constant crossing-point forecasts (see the discussion on equitability in Section 2.2).

3.3. Sensitivity to ensemble size and climatology definition

We focus now on the case where the forecast F and the climatology G are empirical distribution functions. For

example, F can be derived from an ensemble forecast and G based on a set of quantiles. Illustrations in Figs. 4 and 5 are based on an ensemble forecast with 50 members and a climatology defined by 99 quantile levels (1%, 2%, ..., 98%, 99%). We recall that the ensemble size is denoted M and the number of quantiles is denoted n_q .

The score sensitivity to M and n_q is analysed using both the score S and the diagonal score algorithm in Appendix E. Figure 7 shows the diagonal score as a function of the size of the ensemble forecast for three different climate representations defined by equidistant quantiles on $[0, 1]$ with intervals $\frac{1}{n_q}$, n_q taking value 5, 10 and 50. Results obtained with S are shown only for $n_q = 50$ for the sake of the plot readability. More precisely, Fig. 7a,b compare results for 2m temperature and daily precipitation ENS forecasts (European domain, day 5 in the lead time, Summer 2018), respectively. The respective scores when $M = 50$ are used as reference. As a consequence, all curves converge to 1 for $M = 50$.

The ensemble size is a critical parameter in the design of an ensemble system (Leutbecher, 2019). Figure 7 shows the positive impact of increasing the ensemble size on the forecast performance. No qualitative differences appear between the results obtained with S and the ones obtained with the diagonal score. Comparing 2m temperature and daily precipitation plots, the ensemble size has a smaller impact on the scores in the former case. The score converges also more rapidly with increasing M values for 2m temperature forecasts. In addition, we note that more quantile levels in the climate definition (e.g. $n_q = 50$ rather than $n_q = 5$) allows a finer estimation of the ensemble size effect on the score. Figure 7 clearly illustrates that results might differ as a function of the score computation approach and setup, i.e. the number n_q of climate quantile levels itself. Therefore, it is important to communicate this information along with the forecast performance results.

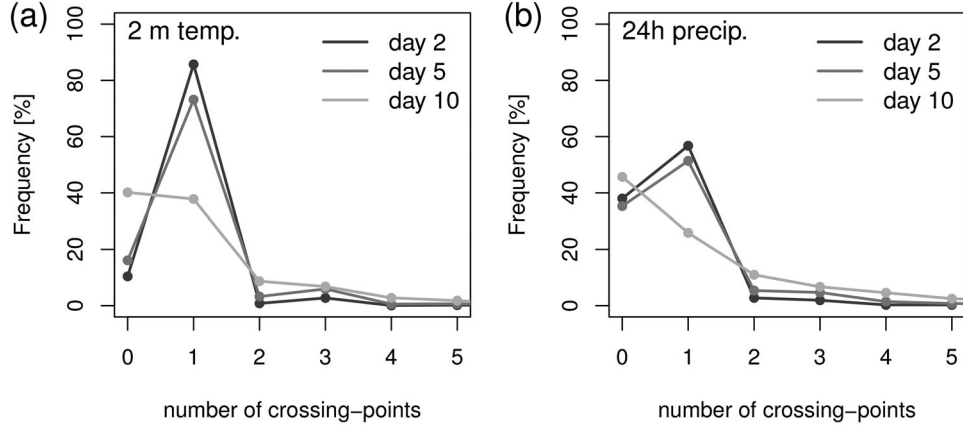


Fig. 6. Distribution of intersection points between each pair of forecast and climate probability distributions. Results for (a) 2 m temperature, and (b) daily precipitation at day 2, 5, and 10, in July 2018, at station level over Europe.

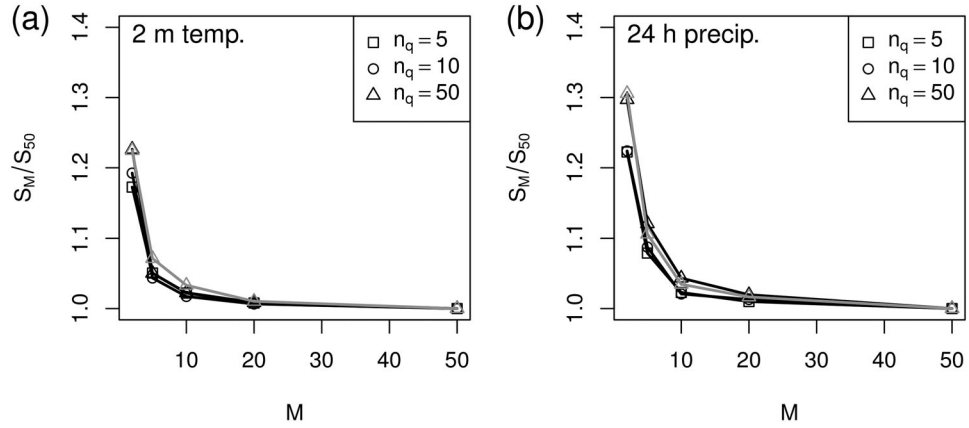


Fig. 7. Score sensitivity to ensemble size and climatology definition: score as a function of the number of ensemble members M (S_M) relative to the score when $M=50$ (S_{50}) for different climate definitions. Results for scores computed with the diagonal score are shown in black, results based on Eq. (3) in grey, for 2m temperature (a) and daily precipitation (b).

4. Conclusion

The crossing-point forecast is defined by the intersection point between a forecast cumulative distribution and the corresponding climatology. The crossing-point is a summary of a probabilistic forecast into a single number conveying information about the worst-case scenario which is more likely in the forecast than in the climatology. Is the predicted chance of suffering a loss, due to the occurrence of an exceedance event, higher than that event's climatological frequency? The crossing-point forecast indicates the limit case for which the answer is positive. In weather forecasting, this type of information could be highly relevant for vulnerable users and more generally for users with interest for high-impact events. Further investigations on the crossing-point forecast interpretation and potential applications are encouraged.

A scoring function consistent for the crossing-point forecast exists. A simple error function that applies to the forecast and observed crossing-points is formulated. The resulting score is proper and equitable which makes its application appealing for the comparison of competing forecasts. The link with other scores and concept is also highlighted. The proposed score is equivalent to the diagonal score in the case of the single intersection condition (when a unique forecast crossing-point exists). Moreover, this work helps generalising the concept of 'score in probability space' to the context of ensemble forecast verification.

In practice, one can encounter situations where multiple crossing-points coexist in a single forecast or where forecast and climate distributions (partly) overlay. Such situations are common in ensemble weather forecasting.

Suggestions on how to tackle such practical challenges are provided. In addition, the analysis of the score sensitivity to ensemble size and climatology definition (the number of used climate quantiles) indicates the clear benefits of a finer representation of both the ensemble and the climate distributions. Finally, an algorithm is provided in order to foster verification applications based on the diagonal score. A systematic comparison with other proper scoring rules for the evaluation and ranking of ensemble forecasts is also encouraged as future work.

Acknowledgements

The author is very grateful to Tobias Fissler for inspiring discussions and exchanges on the concept of score consistency, to Martin Janousek for his help designing the diagonal score algorithm, to David Richardson, Martin Leutbecher, and Linus Magnusson for constructive comments on an earlier version of the manuscript. Valuable comments from one anonymous reviewer are also acknowledged.

References

- Ben Bouallègue, Z., Haiden, T. and Richardson, D. S. 2018. The diagonal score: definition, properties, and interpretations. *Q. J. R. Meteorol. Soc.* **144**, 1463–1473. doi:10.1002/qj.3293
- Ehm, W., Gneiting, T., Jordan, A. and Krueger, F. 2016. Of quantiles and expectiles: consistent scoring functions, choquet representations, and forecast rankings. *J. R. Stat. Soc. B* **78**, 1–29.
- Fissler, T., Hlavínová, J. and Rudloff, B. 2019. Elicitability and identifiability of systemic risk measures. *Papers 1907.01306*, arXiv.org.
- Gandin, L. and Murphy, A. H. 1992. Equitable skill scores for categorical forecasts. *Monthly Weather Rev.* **120**, 361–370. doi:10.1175/1520-0493(1992)120<0361:ESSFCF>2.0.CO;2
- Gneiting, T. 2011. Making and evaluating point forecasts. *J. Am. Stat. Assoc.* **106**, 746–762. doi:10.1198/jasa.2011.r10138
- Gneiting, T. and Ranjan, R. 2013. Combining predictive distributions. *Electron. J. Stat.* **7**, 1747–1782.
- Hawkins, D. and Kochar, S. 1991. Inference for the crossing point of two continuous cdf's. *Ann. Stat.* **19**, 1626–1638.
- Lerch, S., Thorarindottir, T. L., Ravazzolo, F. and Gneiting, T. 2017. Forecaster's dilemma: extreme events and forecast evaluation. *Stat. Sci.* **32**, 106–127.
- Leutbecher, M. 2019. Ensemble size: how suboptimal is less than infinity? *Q. J. R. Meteorol. Soc.* **145**, 107–128. doi:10.1002/qj.3387
- Mitchell, K. and Ferro, C. A. T. 2017. Proper scoring rules for interval probabilistic forecasts. *Q. J. R. Meteorol. Soc.* **143**, 1597–1607. doi:10.1002/qj.3029
- Potts, J., Folland, C., Jolliffe, I. and Sexton, D. 1996. Revised LEPS scores for assessing climate model simulations and

long-range forecasts. *J. Climate* **9**, 34–53. doi:10.1175/1520-0442(1996)009<0034:RSFACM>2.0.CO;2

- Prates, F. and Buizza, R. 2011. Pret, the probability of return: a new probabilistic product based on generalized extreme-value theory. *Q. J. R. Meteorol. Soc.* **137**, 521–537. doi:10.1002/qj.759
- Rodwell, M., Richardson, D., Hewson, T. and Haiden, T. 2010. A new equitable score suitable for verifying precipitation in numerical weather prediction. *Q. J. R. Meteorol. Soc.* **136**, 1344–1363. doi:10.1002/qj.656
- Ward, M. N. and Folland, C. K. 1991. Prediction of seasonal rainfall in the north nordeste of brazil using eigenvectors of sea-surface temperature. *Int. J. Climatol.* **11**, 711–743. doi:10.1002/joc.3370110703

Appendix A: Equitability

Consider a constant probabilistic forecast: there is a single and unique crossing-point τ_c for all verifications τ_y . In that case, a score is called equitable if its expected value is independent of τ_c . Noting that by definition τ_y follows a uniform distribution, the expected score $\mathbb{E}_G[S]$ with respect to the observation unconditional distribution G is developed as:

$$\begin{aligned} \mathbb{E}_G[S(\tau_c, \tau_y)] &= \int_0^{\tau_c} ((1 - \tau_y)^2 - (1 - \tau_c)^2) d\tau_y + \int_{\tau_c}^1 (\tau_y^2 - \tau_c^2) d\tau_y \quad (\text{A1}) \end{aligned}$$

$$= \left(\tau_y - \tau_y^2 + \frac{1}{3} \tau_y^3 - (1 - \tau_c)^2 \tau_y \right) \Big|_0^{\tau_c} + \left(\frac{1}{3} \tau_y^3 - \tau_c^2 \tau_y \right) \Big|_{\tau_c}^1 \quad (\text{A2})$$

$$= \tau_c - \tau_c^2 + \frac{1}{3} \tau_c^3 - (1 - \tau_c)^2 \tau_c + \frac{1}{3} - \tau_c^2 - \frac{1}{3} \tau_c^3 + \tau_c^3 \quad (\text{A3})$$

$$= \frac{1}{3}. \quad (\text{A4})$$

Appendix B: The diagonal score revisited

We show here that the diagonal score can be interpreted as a measure of forecast performance in ‘probability space’. For this purpose, we recall the definition of the elementary score s introduced by Ehm et al. (2016). Denoting x a point forecast issued when the observation y realizes, the scoring function s is defined as:

$$s(x, y) = \begin{cases} \tau & \text{if } y > \theta \geq x \\ 1 - \tau & \text{if } x > \theta \geq y \\ 0 & \text{otherwise} \end{cases} \quad (\text{B1})$$

where $\theta \in \mathbb{R}$ is a threshold defining an event and $\tau \in (0, 1)$ is the score penalty parameter. The scoring function s is consistent for the quantile functional at quantile level τ denoted q_τ .

The peculiarity of the so-called *diagonal elementary score* is that the relationship between the penalty τ and

the threshold θ is fixed such that $\theta = G^{-1}(\tau)$ with G the climatology probability distribution:

$$d_G(q_\tau, y) = \begin{cases} \tau & \text{if } y > G^{-1}(\tau) \geq q_\tau \\ 1 - \tau & \text{if } q_\tau > G^{-1}(\tau) \geq y \\ 0 & \text{otherwise} \end{cases} \quad (\text{B2})$$

With $p_0 = 1 - \tau$, the climatological frequency of the event defined by θ , Eq. (B2) can be derived from the error matrix in Table 1.

Using the following notations:

- τ_y the quantile level such that $y = G^{-1}(\tau_y)$,
- τ_f the quantile level such that $F^{-1}(\tau_f) = G^{-1}(\tau_f)$

with F the forecast probability distributions and given that a unique crossing-point forecast exists, and starting from the definition in Eq. (20) in the work by Ben Bouallègue et al. (2018), the diagonal score is developed as follows:

$$D_G(F, y) = \int_0^1 d_G(q_\tau, y) d\tau \quad (\text{B3})$$

$$= \int_0^1 (\tau \mathbb{I}[y \geq G^{-1}(\tau) > F^{-1}(\tau)] + (1 - \tau) \mathbb{I}[F^{-1}(\tau) \geq G^{-1}(\tau) > y]) d\tau \quad (\text{B4})$$

$$= \mathbb{I}[\tau_y \geq \tau_f] \int_{\tau_y}^{\tau_f} \tau d\tau + \mathbb{I}[\tau_y < \tau_f] \int_{\tau_y}^{\tau_f} (1 - \tau) d\tau \quad (\text{B5})$$

$$= \frac{1}{2} \mathbb{I}[\tau_y \geq \tau_f] (\tau_y^2 - \tau_f^2) + \frac{1}{2} \mathbb{I}[\tau_y < \tau_f] ((1 - \tau_y^2) - (1 - \tau_f^2)) \quad (\text{B6})$$

$$= \frac{1}{2} S(\tau_f, \tau_y). \quad (\text{B7})$$

Appendix C: Interval probabilistic forecasts

Consider the probability forecast $p_f \in [0, 1]$ for a binary event, and p_0 the climatological probability of occurrence of this event. We are interested in whether $p_f > p_0$ or not. So p_f is transformed into an interval probabilistic forecast on two possible ranges. The two intervals are denoted $I_0 := [0, p_0]$ and $I_1 := (p_0, 1]$. The interval probabilistic forecast is I_0 if $p_f \in I_0$, I_1 otherwise.

Following Eq. (7) in the study by Mitchell and Ferro (2017), a proper score s for interval probabilistic forecasts has the following form for each $k = 1, \dots, n - 1$, with n the number of intervals:

$$\begin{aligned} s(I_k, 0) - s(I_{k+1}, 0) &= -a_k \gamma_k \\ s(I_k, 1) - s(I_{k+1}, 1) &= (1 - a_k) \gamma_k \end{aligned} \quad (\text{C1})$$

with γ_k a non-negative constant, and where $s(I_{[1]}, 0)$ and $s(I_{[1]}, 1)$ are the penalties associated with the forecast interval $I_{[1]}$ and the occurrence and non-occurrence of the

event, respectively. The parameters a_k define the intervals on $[0, 1]$ with $0 = a_0 < a_1 < \dots < a_n = 1$.

In the case where $n=2$ with the intervals definition implying that $a_1 = p_0$, and setting the constant $\gamma_1 = 1$, Eq. (C1) becomes:

$$\begin{aligned} s(I_0, 0) - s(I_1, 0) &= -p_0 \\ s(I_0, 1) - s(I_1, 1) &= (1 - p_0) \end{aligned} \quad (\text{C2})$$

Considering no penalty for correct forecasts, correct negative ($s(I_0, 0) = 0$) or correct positive ($s(I_1, 1) = 0$) cases, we obtain $s(I_1, 0) = p_0$ and $s(I_0, 1) = 1 - p_0$, that is the error matrix in Table 1.

Appendix D: Scoring function consistency

A statistical functional is a mapping from a class of probability distributions $\mathcal{F}$ to the power set of $\mathbb{R}$, $T : \mathcal{F} \rightarrow 2^{\mathbb{R}}$. A scoring function $S : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ is $\mathcal{F}$ -consistent for T if

$$E_{Y \sim F}[S(t, Y)] \leq E_{Y \sim F}[S(x, Y)] \quad (\text{D1})$$

for all $F \in \mathcal{F}$, for all $t \in T(F)$, and for all $x \in \mathbb{R}$. It is strictly $\mathcal{F}$ -consistent for T if equality implies that $x \in T(F)$.

Similarly, a scoring rule is a map $R : \mathcal{F} \times \mathbb{R} \rightarrow \mathbb{R}$, tailored to evaluate probabilistic forecasts. It is proper on $\mathcal{F}$ if

$$E_{Y \sim F}[R(F, Y)] \leq E_{Y \sim F}[R(G, Y)] \quad (\text{D2})$$

for all $F, G \in \mathcal{F}$. It is strictly proper on $\mathcal{F}$ if equality implies that $G = F$. Recall that any (strictly) $\mathcal{F}$ -consistent scoring function induces a proper scoring rule on $\mathcal{F}$ (Theorem 3 in Gneiting, 2011) in that $R(F, y) = S(t_F, y)$ for some $t_F \in T(F)$.

Let G be some probability distribution function. We argue that $S_G : [0, 1] \times \mathbb{R} \rightarrow \mathbb{R}$, $S_G(\tau, y)$ defined in Eq. (3), is an $\mathcal{F}$ -consistent score for the crossing-point functional $T_G : \mathcal{F} \rightarrow 2^{[0, 1]}$, $T_G(F) = \{z \in [0, 1] : \exists x \in \mathbb{R} \text{ such that } F(x) = G(x) = z\}$. Indeed, Appendix B shows that the diagonal score corresponds to half the scoring rule induced by S_G . Therefore, the reverse implication of Theorem 3 in the study by Gneiting (2011) implies the $\mathcal{F}$ -consistency of S_G for T_G .

We give a concise interpretation of the crossing-point functional on the level of the prediction space setting (as defined by Gneiting and Ranjan, 2013). Fix some random variable Y with unconditional (climatological) distribution G . Suppose a forecaster has some information $\mathcal{A}$ (mathematically speaking $\mathcal{A}$ is a σ -algebra). Then the ideal probabilistic forecast is the conditional distribution of Y given $\mathcal{A}$, $F = \mathcal{L}(Y|\mathcal{A})$. Accordingly, the ideal crossing-point forecast, given $\mathcal{A}$, is any point in $T_G(F)$. Note that if $\mathcal{A}$ does not contain any relevant information about Y (i.e. Y and $\mathcal{A}$ are independent), and in particular if $\mathcal{A}$ does not

have any information at all (so it is trivial), then the ideal probabilistic forecast is G itself, the unconditional distribution. Accordingly, any best uninformed crossing-point forecast is a number $\tau \in T_G(G) = [0, 1]$. Therefore, the consistency of S_G directly recovers the equitability result of [Appendix A](#).

Appendix E: Python algorithm

We provide here an easy-to-read python algorithm written with the assessment of ensemble forecasts of weather variables in mind, considering non-decreasing quantiles, excluding non-unique quantiles, and recommending climatological quantile levels uniformly spaced in the unit interval:

```
import numpy as np
```

```
def compute_ds(qclim, tau_clim, obs, ens):
```

```
    """Computes the diagonal score as
    the average diagonal elementary score
    over all unique climate quantile
    levels.
```

```
    Parameters:
```

```
        qclim (np.array): array of equidistant
        increasing quantiles representing
        the climatology
```

```
        tau_clim (np.array): array of quantile
        levels corresponding to qclim
```

```
        obs (float): observation value
```

```
        ens(np.array): array with the ensemble
```

```
        members
```

```
    Returns:
```

```
        ds (float): the score value
```

```
    """
```

```
    mask = []
    nq = len(qclim)
    perc_ref = -np.inf
    for iq in range(nq):
        pos_mask = qclim[iq] > perc_ref
        perc_ref = qclim[iq]
        mask.append(pos_mask)
    perc_ref = np.inf
    for iq in range(nq)[::-1]:
        pos_mask = qclim[iq] < perc_ref
        perc_ref = qclim[iq]
        mask[iq] &= pos_mask
    nc = 0.
    dse = 0.
    for iq in range(nq):
        tau = tau_clim[iq]
        obs_ev = obs > qclim[iq]
        pre_ev = ens > qclim[iq]
        p = pre_ev.mean().astype(float)
        dst = obs_ev*(p <= (1.-tau))*tau + \
            (1.-obs_ev)*(p > (1.-tau))*(1.-tau)
        dse += dst*mask[iq]
        nc += mask[iq]
    ds = 2.*dse/nc
    return ds
```