RMetS
Royal Meteorological Society

Predicting outliers in ensemble forecasts

Stefan Siegert,[a]* Jochen Bröcker[a,b] and Holger Kantz[a]
[a]*Max-Planck-Institut für Physik komplexer Systeme, Dresden, Germany*
[b]*Centre for the Analysis of Time Series, London School of Economics, London, UK*

*Correspondence to: S. Siegert, Max-Planck-Institut, Nöthnitzer Strasse 38, 01187 Dresden, Germany.
E-mail: siegert@pks.mpg.de

An ensemble forecast is a collection of runs of a numerical dynamical model, initialized with perturbed initial conditions. In modern weather prediction for example, ensembles are used to retrieve probabilistic information about future weather conditions. In this contribution, we are concerned with ensemble forecasts of a scalar quantity (say, the temperature at a specific location). We consider the event that the verification is smaller than the smallest, or larger than the largest ensemble member. We call these events outliers. If a K-member ensemble accurately reflected the variability of the verification, outliers should occur with a base rate of $2/(K+1)$. In operational forecast ensembles though, this frequency is often found to be higher. We study the predictability of outliers and find that, exploiting information available from the ensemble, forecast probabilities for outlier events can be calculated which are more skilful than the unconditional base rate. We prove this analytically for statistically consistent forecast ensembles. Further, the analytical results are compared to the predictability of outliers in an operational forecast ensemble by means of model output statistics. We find the analytical and empirical results to agree both qualitatively and quantitatively. Copyright © 2011 Royal Meteorological Society

Key Words: forecast probability; predictability; Talagrand diagrams; ROC curves; skill scores

Received 21 December 2010; Revised 27 April 2011; Accepted 23 May 2011; Published online in Wiley Online Library 19 July 2011

Citation: Siegert S, Bröcker J, Kantz H. 2011. Predicting outliers in ensemble forecasts. *Q. J. R. Meteorol. Soc.* **137**: 1887–1897. DOI:10.1002/qj.868

1. Introduction

Ensemble weather forecasts are used operationally to issue probabilistic predictions of future atmospheric conditions (Leith, 1974; Sivillo *et al.*, 1997; Kalnay, 2003). For an ensemble forecast, a collection (an ensemble) of initial conditions is evolved according to the prognostic equations of an atmospheric model. Model parameters might also differ between ensemble members. All members of a forecast ensemble verify at the same time. A working hypothesis as to the interpretation of ensembles is that each initial condition is an equally likely estimate of the present state of the atmosphere. This is assumed despite the fact that the initial conditions are by construction neither independent nor a random sample from the full state space of the model (Descamps and Talagrand, 2007). Under the working hypothesis, the resulting ensemble of predictions comprises a collection of equally likely scenarios of the future state of the atmosphere, estimated by the numerical model. The quality of the ensemble forecast should ultimately be evaluated by comparison to the actual verifying observation, or verification for short. In particular, the working hypothesis implies that ensemble members and verification can be considered draws from the same distribution, the forecast distribution. This is encapsulated in the notion of statistical consistency (Anderson, 1997; Wilks, 2006).

In operational weather forecasts, ensembles are often not consistent. A common problem is lack of ensemble dispersion. The typical difference between ensemble members is smaller than the typical difference between verification and ensemble members. Underdispersiveness causes overconfident uncertainty estimates. If the complete state vector is projected onto a single variable in a single location, ensemble members and the verification become

scalar quantities. Underdispersiveness then leads to regular occurrence of events where the verification falls outside the range of the ensemble. These events, which we call outliers, are the main subject of this study. Outliers fall into either one of the tails of the forecast distribution constructed from the ensemble, that is into regions to which low probability has been assigned. This leads to the occurrence of unexpected events–events deemed improbable by the forecast ensemble and whose occurrence might be detrimental if decisions are based on the forecast.

The relevance of outliers in evaluating ensemble forecasts has been acknowledged by different authors. In Buizza and Palmer (1998), the outlier statistic is defined as the percentage of outliers over a prescribed time interval, where outliers are defined as analyses lying outside the ensemble range. The analysis is defined as the projection of the observations into the numerical model by means of data assimilation (Kalnay, 2003). It is argued that the percentage of outliers in a consistent K-member ensemble should not significantly exceed $2/(K+1) \times 100\%$. The outlier statistic has been used as a performance measure in a comparative study of ensemble prediction systems (Ziehmann, 2000; Buizza *et al.*, 2005). Outliers tend to occur too frequently in operational ensembles, thus rendering these ensembles statistically inconsistent. The choice of the initial conditions affects the percentage of outliers (Barkmeijer *et al.*, 1999, Bowler, 2006).

One could naïvely assume that the probability of occurrence of outliers is independent of the present state of the numerical model, or of the ensemble. This would imply that the best possible prediction of such outliers is given by their unconditional base rate of occurrence. We challenge this view and consider the following question: Can outliers in ensemble forecasts be predicted and, if so, how much better can we do than just using the unconditional base rate for prediction?

In section 2 we show that outliers are predictable beyond the base rate in consistent forecast ensembles. Analytical results for performance of outlier prediction in consistent ensembles are discussed using standard performance indices. We compare our analytical results with findings from an experiment where we investigate the predictability of outliers in a temperature forecast ensemble for Dresden, Germany. In section 3, we show that there is potential predictability of outliers in this operational (not necessarily consistent) ensemble by investigating variations in conditional outlier frequency. In section 4, we make use of this knowledge and issue predictions in the operational ensemble based on statistical models. Performance measures so obtained are contrasted with the consistent case. Section 5 provides conclusions.

2. Outliers in a consistent forecast ensemble

A forecast ensemble is statistically consistent if all ensemble members as well as the corresponding verification can be considered as independent random draws from an underlying forecast distribution. For the investigation of outliers, we consider the case where the ensemble members are scalar quantities. In this situation, mathematically speaking, we can always assume that the ensemble and the verification arise as follows. Let $\Gamma(\cdot) : \mathbb{R} \to [0,1]$ be the forecast distribution, written here as a cumulative distribution function. The ensemble members can be considered as K random samples $e_1, \ldots, e_K$ drawn from Γ. The verification η is just another draw from Γ. We will write

$$E \equiv (e^{[1]}, \ldots, e^{[K]}) \text{ with } -\infty < e^{[1]} \leq \ldots \leq e^{[K]} < \infty$$

for the ensemble members ordered by increasing magnitude. This model appropriately describes a consistent ensemble forecast of a scalar quantity in the most general way. In this section, we will assume that the forecast distribution Γ is known. This assumption is typically not valid in realistic forecast settings. Based on this assumption, however, we are able to compute several benchmark results concerning the performance of outlier prediction.

The probability σ for the event 'η is an outlier', conditional on the ensemble and on Γ, is given by the mass of probability concentrated outside the ensemble, namely

$$\begin{aligned} \sigma &\equiv \Pr\left(\left\{\eta \leq e^{[1]}\right\} \cup \left\{\eta > e^{[K]}\right\} \mid E, \Gamma\right) \\ &= \Gamma\left(e^{[1]}\right) + 1 - \Gamma\left(e^{[K]}\right). \end{aligned} \tag{1}$$

Obviously, the outlier probability σ in a consistent ensemble is not always equal to $2/(K+1)$, but it is a fluctuating quantity. We show in Appendix A that σ is distributed with density

$$\begin{aligned} f_\sigma(\tau) &\equiv p(\sigma = \tau) \\ &= K(K-1)(1-\tau)^{K-2}\tau. \end{aligned} \tag{2}$$

It is remarkable that, even though σ itself is dependent on Γ as Eq. (1) demonstrates, its distribution f_σ does not depend on Γ. The only parameter determining the shape of the distribution is the ensemble size K. From Eq. (2) we obtain the intuitively obvious result that the unconditional base rate (which is equal to the expectation value of the outlier probability) in a consistent ensemble is given by

$$\mathbb{E}[\sigma] = \int_0^1 \mathrm{d}\tau\, f_\sigma(\tau)\,\tau = \frac{2}{K+1}. \tag{3}$$

Distributions of outlier probabilities for different ensemble sizes are shown in Figure 1(a). Larger ensemble sizes lead to smaller base rates, roughly indicated by the location of the maxima of the distributions, and less variability, as indicated by their spread.

Based on Eq. (2), it is possible to calculate expectation values of performance measures for the prediction of outliers in consistent forecast ensembles. Performance measures which are widely used for the evaluation of probabilistic forecasts of binary events include the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC; Egan, 1975) as well as proper scoring rules (Gneiting and Raftery, 2007). The ROC curve corresponding to σ is a plot of the hit rate $H(\zeta) = \Pr(\sigma > \zeta \mid y = 1)$ over the false alarm rate $F(\zeta) = \Pr(\sigma > \zeta \mid y = 0)$, where ζ is a varying threshold that parametrizes the ROC curve. In our case, the event indicator y equals unity if an outlier occurs and zero otherwise. The AUC is then given by $\int_0^1 \mathrm{d}F\, H(F)$. A ROC curve that lies significantly above the diagonal, or equivalently, an AUC significantly exceeding $1/2$, indicates skilful discrimination between events and non-events.

A scoring rule is a function $s(\sigma, y)$ that assigns to a forecast-verification pair (σ, y) a real number which quantifies how

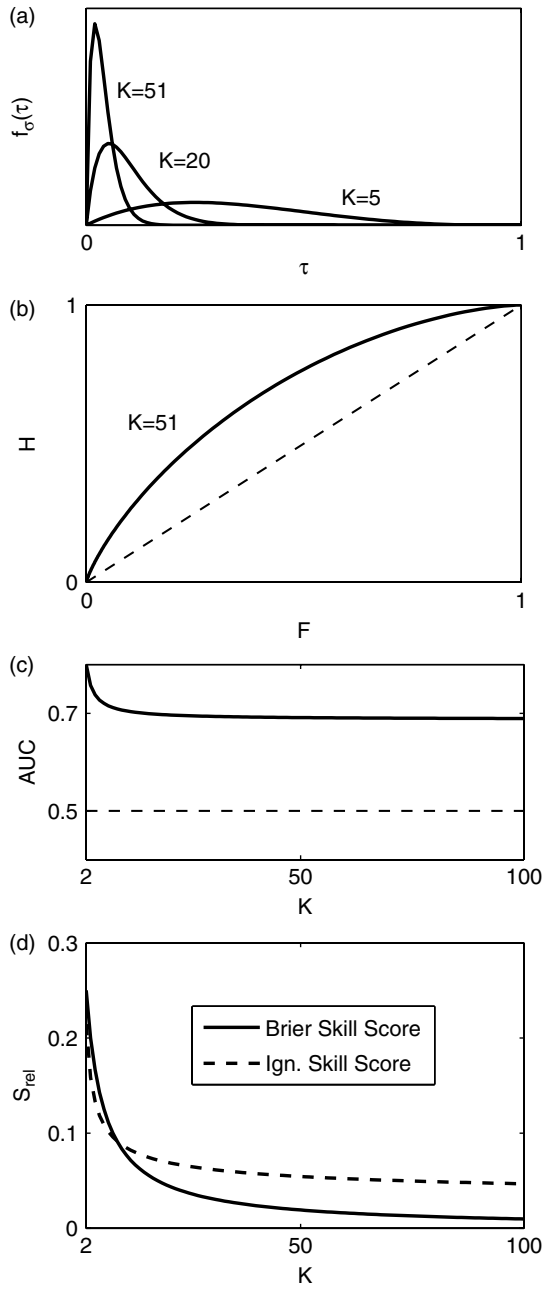

Figure 1. (a) Distributions of outlier probabilities σ for different values of the ensemble size K. (b) Receiver operating characteristic (ROC) curves with hit rate H plotted versus false-alarm rate F for the probabilistic forecast σ for $K = 51$ (solid), and for the constant base-rate forecast (dashed). (c) Area under the curve (AUC; $\int_0^1 H(F)\,\mathrm{d}F$) plotted versus K for σ (solid) and for the constant base-rate forecast (dashed). (d) Brier Skill Score and Ignorance Skill Score (improvement over base-rate forecast) plotted versus K. These results hold for any ensemble which is statistically consistent. All performance measures indicate significantly better prediction success than the base-rate forecast. Note that the dependence on the ensemble size is more pronounced for skill scores than for AUC.

well σ performed at predicting y. A scoring rule is proper if its expected value is maximized by assigning the probability p to an event whose true probability of occurrence is indeed p, that is by reporting probabilities honestly. Two widely used proper scoring rules are the Brier Score (Brier, 1950)

$$s(\sigma, y) = (y - \sigma)^2 = y(1 - \sigma)^2 + (1 - y)\sigma^2, \qquad (4)$$

and the Ignorance Score (Roulston and Smith, 2002)

$$s(\sigma, y) = -y \log_2(\sigma) - (1 - y) \log_2(1 - \sigma). \qquad (5)$$

Both scoring rules assign lower numbers to better forecasts and become zero if the forecast is perfect. A skill score S_{rel} (the subscript rel denotes 'relative') is given by the relative improvement of the expected score $\mathbb{E}[s]$ of the forecast under consideration over the expected score $\mathbb{E}[s_{\mathrm{ref}}]$ of a reference forecast (Wilks, 2006). Thus, skill scores have the form

$$S_{\mathrm{rel}} = 1 - \frac{\mathbb{E}[s]}{\mathbb{E}[s_{\mathrm{ref}}]}, \qquad (6)$$

if the scoring rule is zero for a perfect forecast. A skill score defined by Eq. (6) is positively oriented. It is equal to unity if the forecast that is evaluated is perfect, and equal to zero if the forecast is as good as the reference forecast.

The calculation of performance measures of outlier prediction in consistent forecast ensembles will be illustrated using the Brier Score. We first calculate the expected Brier Score for a given value of σ (say $\sigma = \tau$) by noting that

$$\begin{aligned}\mathbb{E}\big[(y - \sigma)^2 \mid \sigma = \tau\big] &= \Pr(y = 1 \mid \sigma = \tau)(1 - \tau)^2 \\ &\quad + \Pr(y = 0 \mid \sigma = \tau)\tau^2 \\ &= \tau(1 - \tau)^2 + (1 - \tau)\tau^2 \\ &= \tau(1 - \tau). \end{aligned} \qquad (7)$$

Here we made use of the reliability of the forecast σ, that is σ satisfies $\Pr(y = 1 \mid \sigma = \tau) = \tau$ by construction. The expected Brier Score is obtained by weighting the conditional expectation value in Eq. (7) by the frequency of occurrence of the particular value of τ. This frequency is given by $f_\sigma(\tau)$ in Eq. (2). Integrating over all possible values of τ yields the expectation value of the Brier Score,

$$\begin{aligned}\mathbb{E}\big[(y - \sigma)^2\big] &= \int_0^1 \mathrm{d}\tau\, f_\sigma(\tau)\, \tau(1 - \tau) \\ &= \frac{2(K - 1)}{(K + 1)(K + 2)}, \end{aligned} \qquad (8)$$

which follows from integration by parts. Our main question is how much better outliers can be predicted than by their unconditional base rate $2/(K + 1)$. Hence, we use the base-rate forecast as the reference forecast to calculate the Brier Skill Score from Eq. (6). The expected Brier Score for the constant base-rate forecast is obtained by replacing τ by $2/(K + 1)$ in Eq. (7), that is

$$\mathbb{E}[s_{\mathrm{ref}}] = \frac{2(K - 1)}{(K + 1)^2}. \qquad (9)$$

Substituting Eqs (8) and (9) into Eq. (6) yields

$$S_{\mathrm{rel}} = \frac{1}{K + 2} \qquad (10)$$

for the Brier Skill Score. It is worth noting that, once again, Eq. (10) is independent of the forecast distribution Γ.

In Appendix B we derive further expressions for expected hit rate, false alarm rate, AUC and Ignorance Skill Score. Some results are illustrated in Figure 1(b–d). ROC curves (Figure 1(b)), AUCs (Figure 1(c)) and skill scores (Figure 1(d)) indicate that σ has significantly better predictive skill than the base-rate forecast. Note that the K-dependence of the skill scores is more pronounced than that

of the AUC. In fact, it can be shown that Ignorance Skill Score and Brier Skill Score converge to zero for $K \to \infty$ while the AUC converges to 11/16 for $K \to \infty$ (Appendix C). In section 4, the results obtained for the consistent ensemble will be compared with outlier prediction success in an operational forecast ensemble.

3. Stratified Talagrand diagrams

In this section, we follow Bröcker (2008) and test for consistency* under stratification. A forecast stratum is defined as a group of forecast cases which have been selected according to a criterion that is conditional on the forecast distribution. In Bröcker (2008) it is argued that, if a forecast ensemble is consistent, it should therefore be independent of stratification. That means that a consistency test should be passed by the collection of all forecast cases as well as by any stratum of the latter.

We use stratified v_l-diagrams (Bröcker, 2008), a modification of the Talagrand diagram (Anderson, 1996; Hamill and Colucci, 1997; Talagrand *et al.*, 1997) to assess ensemble consistency. The Talagrand diagram is constructed from a collection of ensemble-verification pairs. At each time, the K ensemble members $e_1, \ldots, e_K$ are ordered by increasing magnitude, $e^{[0]} < e^{[1]} \leq \cdots \leq e^{[K]} < e^{[K+1]}$, where $e^{[0]}$ and $e^{[K+1]}$ are set to $-\infty$ and $+\infty$, respectively. The verification η is said to have rank l, if $e^{[l-1]} \leq \eta < e^{[l]}$, which implies that $l \in [1, K+1]$. The Talagrand diagram is the histogram of observed frequencies of verification ranks $1, \ldots, K+1$. It is asymptotically flat for a consistent forecast ensemble.

If constructed from a finite number of samples, the Talagrand diagram is subject to random fluctuations. In the v_l-diagram, deviations from flatness due to finite samples are taken into account. Under the assumption of consistency, the number of cases in which the verification assumes rank l should follow a binomial distribution with parameters $n = N$ and $p = 1/(K+1)$. Here N is the number of ensemble-verification pairs that are tested and K is the number of ensemble members. In the v_l-diagram, the values $v_l = B\{r_l; N, 1/(K+1)\}$ are reported for each rank l, where r_l is the number of occurrences of verification rank l in the test dataset, and $B(x; n, p)$ is the value of the cumulative binomial distribution with parameters n and p, evaluated at x. Thus, instead of showing the number of occurrences of each rank, the likelihood of observing at most that number under the hypothesis of consistency is shown.

In order to reject the consistency hypothesis based on individual values of v_l, we have to correct for multiple testing. The probability of at least one out of $K+1$ values lying outside the, say, 95% confidence interval equals $1 - 0.95^{K+1}$ which is in general larger than 0.05. For that reason, we do not indicate $0.95 \times 100\%$ intervals but $0.95^{1/(K+1)} \times 100\%$ intervals in the v_l-diagrams. This so-called Bonferroni correction implies that the consistency hypothesis is rejected at confidence level 0.95 if at least one of the v_l values falls outside the confidence interval.

In the present study we evaluate the forecast ensemble provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). We focus on 2 m temperature forecasts that were issued for Dresden/Germany (WMO number 10488) during the period 2001–2006 (inclusive). All forecasts verify at noon. We carried out simple post-processing by correcting for constant and seasonal bias of the ensemble mean. In the following, these data will be referred to as 'the operational ensemble'.

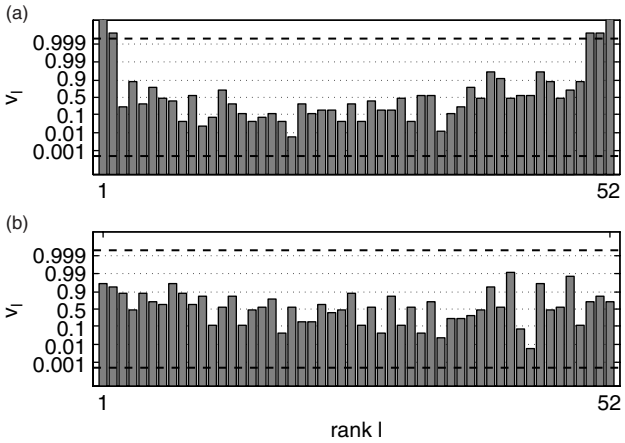

Figure 2. The v_l-diagram for the 2 m temperature forecast at WMO 10488, lead time 144 h, stratified by ensemble range $e^{[K]} - e^{[1]}$. (a) uses that half of the cases in which the ensemble has the lowest range, and (b) uses ensembles with the highest range. The region between the bold dashed lines corresponds to the 95% confidence interval corrected for multiple sampling. If the ensemble range is low, outliers occur too frequently so that the histogram cannot be considered flat.

Both Figures 2 and 3 provide examples of conditional variations of the occurrence of outliers in the operational ensemble. In Figure 2, forecast cases were stratified by ensemble range $e^{[K]} - e^{[1]}$. Two separate v_l-diagrams were constructed after splitting the collection of all forecast-verification pairs into the 50% for which the ensemble has lowest range and those for which it has highest range. Bold dashed lines indicate the Bonferroni-corrected 95% confidence interval.

In low-range ensembles, the v_l-diagram is significantly U-shaped (Figure 2(a)). Outliers occur frequently enough so that the U-shape of the Talagrand diagram cannot be attributed to finite sample effects. On the other hand, forecast cases in the high-range stratum cannot be considered inconsistent based on the v_l-diagram. In the present study, we are not really interested in proving or disproving the consistency of this particular forecast ensemble. Rather, it is our goal to identify factors that favour the occurrence of outliers in forecast ensembles. The ensemble range clearly is indicative of the occurrence of outliers, as shown in Figure 2.

As an interesting aside, we note that even a perfectly consistent ensemble would show the behaviour observed in Figure 2. Outliers occur more frequently in the low-range stratum than in the high-range stratum and this is not a consequence of the inconsistency of that forecast ensemble. Rather it points out the fact that a low range of the forecast ensemble does not necessarily imply increased certainty in the future state of the atmosphere. A low range of a finite sample from any distribution can occur by chance, thus increasing the probability of an outlier on that particular forecast case. For this reason, ensembles with below-average range will have above-average probability of leading to an outlier. Consequently, an above-average fraction of outliers will occur on low-range cases. We conclude that the

*In Bröcker (2008), the term 'reliability' is used interchangeably with what we call consistency in the present study. We distinguish between the two and reserve the term reliability for probabilities p in binary forecast problems that satisfy the condition $\Pr(y = 1 \mid p) = p$.

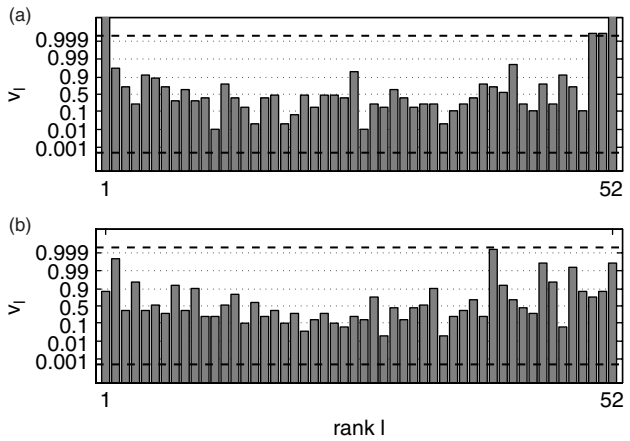

Figure 3. As Figure 2, but stratified by season, using forecast-verification pairs (a) from December to May (inclusive), and (b) from June to November. Outliers occur too frequently in (a), which is not the case for (b).

ensemble range is not a suitable criterion for the assessment of consistency by means of forecast stratification.

Similar behaviour to Figure 2 can be observed if the forecast cases are stratified by forecast time, as in Figure 3. During the period December to May (inclusive), outliers occur significantly more often than would be expected if the rank counts in the corresponding Talagrand diagram were binomially distributed. For the second period, the consistency hypothesis cannot be rejected based on the v_l-diagram. We have identified a further factor, namely the current season, which is indicative of the occurrence of outliers.

As a further aside, it is not clear if stratification with respect to season in order to assess ensemble consistency is justified from a theoretical point of view. Stratification with respect to season seems appropriate and Figure 3 seems to indicate seasonal inconsistency of the operational ensemble. However, the season is not explicitly a function of the forecast distribution. It is only implicitly introduced in the form of changing solar forcing in the numerical model. We stratify with respect to a criterion which cannot directly be inferred from the ensemble, but only indirectly from the current temperature values. In a mean-sea-level pressure ensemble, such an inference of the season based on the current state of the ensemble would be even harder. As stated at the beginning of section 3, a fair stratum should depend only on information that is directly available from the ensemble at the forecast time. Hence, similarly to Figure 2, Figure 3 provides a means of distinguishing conditions that favour outliers but it does not necessarily prove inconsistency of the operational ensemble.

4. Predicting outliers in an operational ensemble

We have seen in Figures 2 and 3 in the previous section that ensemble range and season might serve as indicators of the frequency of outliers in the operational forecast ensemble. In the present section, we shall exploit this finding and estimate outlier probabilities in the operational ensemble based on information that is available at forecast time. This means we are going to tune an algorithm that establishes a relation –a model –between the available inputs and the probability of occurrence of an outlier. A well-studied algorithm for this purpose is logistic regression (Park and Hastie, 2008; Hastie *et al.*, 2009).

Let us first outline the modelling procedure and fix notation. Suppose we have at our disposal event indicators $\mathbf{y} = (y_1, \ldots, y_N)^{\mathrm{T}}$, where $y_t = 1$ if the event occurs at time t and $y_t = 0$ otherwise. At each time t we have access to an input vector $\mathbf{x}_t \in \mathbb{R}^M$. This vector contains the information based on which we want to estimate $p_t \equiv \Pr(y_t = 1 \mid \mathbf{x}_t)$, the posterior event probability at time t. In logistic regression, a linear model is fitted to the logit transform of p_t, that is $\mathbf{x}_t^{\mathrm{T}}\beta = \log\{p_t/(1-p_t)\}$. In other words, we are seeking a vector of coefficients, β, from which the event probability at time t can be calculated by

$$p_t = p(\mathbf{x}_t, \beta) = \frac{\exp(\mathbf{x}_t^{\mathrm{T}}\beta)}{1 + \exp(\mathbf{x}_t^{\mathrm{T}}\beta)}. \tag{11}$$

The quality of the probability estimate Eq. (11) must be assessed by how well the event probability p_t performs at predicting y_t. For this purpose, we use the proper scoring rules Brier Score and Ignorance Score defined in section 2, Eqs (4) and (5). The empirical score $S(\beta)$ of a model with coefficients β is given by the average score over N cases:

$$S(\beta) = \frac{1}{N}\sum_{t=1}^{N} s\{p(\mathbf{x}_t, \beta), y_t\}. \tag{12}$$

In order to learn the connection between input vectors and event indicators, we require a dataset of examples –the training data. However, the optimal coefficient vector is not necessarily the one that achieves the lowest score inside the training data. For the purpose of prediction, the optimization algorithm has to find a model that scores well not only inside the training data but also in yet-unknown cases with inputs that are not part of the training data. If the complexity of the model is increased by increasing M (the dimension of the input vector $\mathbf{x}_t$), one can make the model fit the known training data arbitrarily well. Usually such an overfitted model generalizes poorly to unknown cases.

Overfitting can be avoided by limiting the variability of the model coefficients (Hastie *et al.*, 2009). In the present study, this is achieved by L_2-regularization, where during the optimization procedure a penalty proportional to the L_2-norm of the coefficient vector is added to the empirical score, that is

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{S(\beta) + \lambda||\beta||^2\right\}. \tag{13}$$

Equation (13) shows that the optimization has to find a compromise between making the model fit the training data and at the same time keeping the coefficients small. The parameter λ controls the amount of regularization, where larger values of λ lead to smaller model coefficients. The problem of finding $\widehat{\lambda}$, the optimal value of λ, which establishes the best compromise between fitting the training data and generalizing to unknown inputs is approached via leave-one-out cross-validation (Hastie *et al.*, 2009). In leave-one-out cross-validation, N different models are trained, where for the optimization of the jth coefficient vector $\beta_{[-j]}$, the jth input-verification pair is omitted from the training data. The leave-one-out score is the empirical mean over the scores that the jth model achieves when evaluated using

Table I. Input combinations used in the logistic regression model. $\mathbf{e}_t = (e_t^{[1]}, \ldots, e_t^{[K]})$ denotes the ensemble at time t, with members ordered by increasing magnitude, and $\omega = (2\pi/365.2425)\ \mathrm{day}^{-1}$.

Combination	
1	$\mathbf{x}_t = (1, \cos\omega t, \sin\omega t)^{\mathrm{T}}$
2	$\mathbf{x}_t = (1, \cos\omega t, \sin\omega t, \cos 2\omega t, \sin 2\omega t)^{\mathrm{T}}$
3	$\mathbf{x}_t = \left(1, \cos\omega t, \sin\omega t, \cos 2\omega t, \sin 2\omega t, e_t^{[1]}, e_t^{[K]}\right)^{\mathrm{T}}$
4	$\mathbf{x}_t = (1, \cos\omega t, \sin\omega t, \cos 2\omega t, \sin 2\omega t, \mathbf{e}_t)^{\mathrm{T}}$

the jth input-verification pair, i.e. the one that was not used during its optimization:

$$S_{\mathrm{loo}}(\lambda) = \frac{1}{N}\sum_{j=1}^{N} s\left\{p(\mathbf{x}_j, \widehat{\beta}_{[-j]}), y_j\right\}. \qquad (14)$$

The regularization parameter $\widehat{\lambda}$ is obtained by minimizing S_{loo} inside the training dataset. The full procedure of fitting N models for each possible value of λ in order to find the best one is computationally very expensive. However, in Bröcker (2010) an efficient technique is presented that facilitates the estimation of $S_{\mathrm{loo}}(\lambda)$ in logistic regression models. The estimate requires only the results of the single optimization of β using the full training dataset, if the optimization is carried out by means of the Newton–Raphson algorithm.

For the purpose of estimating out-of-sample outlier probabilities in the operational ensemble we apply tenfold cross-validation (Hastie *et al.*, 2009) to the ensemble forecast data. We split the 6-year dataset of ensemble forecast-verification pairs into ten contiguous chunks of equal length. One of the chunks is set aside and the union of the other nine is taken as the training dataset. The tenth chunk is taken as the test dataset. The event indicator y_t is set to unity if an outlier occurs at time t and zero otherwise. As model inputs we evaluate four different combinations which are given in Table I. These inputs enable the model to take into account seasonality of outlier frequency, motivated by the results of section 3 as well as dependence on the values of the ensemble members, motivated by Eq. (1). All inputs except the first are standardized to zero mean and unit variance. Using only the training data, the optimal regularization parameter $\widehat{\lambda}$ is estimated by minimizing $S_{\mathrm{loo}}(\lambda)$ with respect to λ.

Having found $\widehat{\lambda}$, a regularized logistic regression model is fit to the training data set. For the optimization, we use a standard Newton–Raphson algorithm with step size halving if overshooting occurs. The optimal model coefficients are the ones that minimize the empirical Brier Score augmented with the penalty given in Eq. (13). The first component of β, the intercept, is not penalized. The resulting optimal coefficients are used to estimate outlier probabilities using the test dataset. This procedure is repeated ten times, where each chunk of the full dataset is used as the test dataset once. The ten test datasets so obtained are put back together to a 6-year dataset of forecast-verification pairs, in which each forecast is independent of its corresponding verification. The datasets of forecast-verification pairs based on different model inputs will be subject to evaluation in the following two sections.

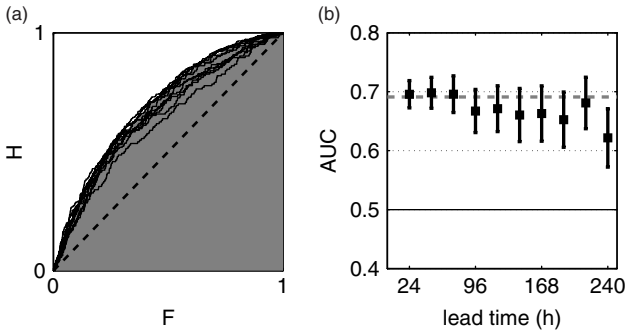

Figure 4. (a) The ROC curves of the logistic regression estimates of p_t in the operational ensemble at different lead times (solid black). The ROC curve for the consistent case ($K = 51$) is highlighted by shading grey the area under that curve. The dashed diagonal line indicates the ROC curve of the constant base-rate forecast; the area under this curve is equal to 1/2. (b) The corresponding AUCs plotted versus lead time. Markers with 95% confidence bars indicate AUCs of p_t in the operational ensemble. The grey dashed line corresponds to the AUC for the consistent case. AUCs at all lead times can be considered significantly larger than 1/2, hence the ROC curves are significantly different from the diagonal. ROC curves and AUCs for outlier prediction in the operational ensemble are close to that of a consistent ensemble for all lead times except 240 h.

4.1. *ROC analysis*

In this section, we treat the estimated outlier probabilities p_t as classifiers and evaluate their performance by means of ROC curves and AUC (Egan, 1975). Since the number of samples of p_t is finite, the hit rate $H(\zeta) = \Pr(p_t > \zeta \mid y_t = 1)$ and the false alarm rate $F(\zeta) = \Pr(p_t > \zeta \mid y_t = 0)$ are estimated by their relative frequencies in the test data. We compute the AUC by trapezoidal approximation. Confidence intervals of the AUC can be constructed based on the assumption of asymptotic normality (DeLong *et al.*, 1988).

In Figure 4, we show ROC curves and AUCs of estimated outlier probabilities for the operational ensemble using input combination 3 from Table I. The ROC curves do not differ much between different lead times except for the one at 240 h. The confidence bands of the AUCs do not overlap 1/2, so we can assume significant skill of our estimates in discriminating outlier events from non-events. Furthermore, the ROC curves of our classifier are close to the theoretical (ideal) ROC curve in a consistent ensemble. Note that for the latter ROC curve it is assumed that the forecast distribution from which the ensemble members are drawn is known at each time and that the verification is known to be another random draw from this distribution. The forecast distribution is, of course, not known in the operational ensemble, but it appears that the ensemble contains enough information to discriminate outliers from non-outliers with equivalent skill as if the forecast distribution were known.

Further, we want to analyse the effect of the individual inputs. Figure 5 depicts differences in AUC between successive input combinations. The error bars indicate 95% confidence regions for the difference in AUC, taking into account correlation of the AUCs caused by the fact that the classifiers under comparison are tested against the same verification data. Furthermore, asymptotic normality of the AUC is assumed. By propagation of error, the variance of the difference between AUC_i and AUC_j is given by

$$\mathrm{var}(\mathrm{AUC}_i - \mathrm{AUC}_j) = \mathrm{var}(\mathrm{AUC}_i) + \mathrm{var}(\mathrm{AUC}_j) - 2\mathrm{cov}(\mathrm{AUC}_i, \mathrm{AUC}_j). \qquad (15)$$

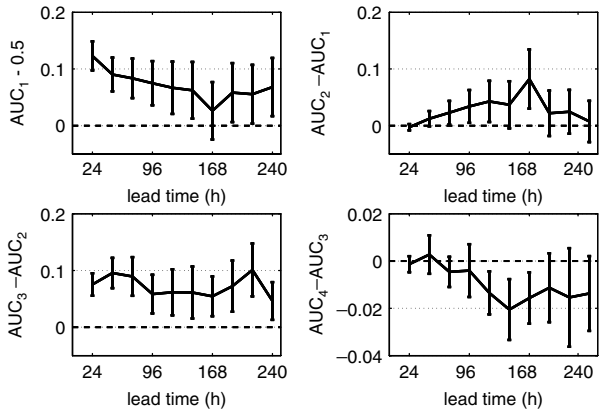

Figure 5. Difference in AUC for estimated outlier probabilities in the operational ensemble obtained for different input combinations (from Table I). AUC_i refers to the AUC obtained by input combination i. Positive values indicate improvement, and error bars are 95% confidence intervals. Seasonality (input combinations 1 and 2) has a significant effect and adds up to 0.1 to the AUC value of 1/2. Combination 2 is especially beneficial at lead time 168 h. The inclusion of the outer ensemble members adds 0.05 to 0.1 to AUC_2. Using the whole ensemble does not add significantly to the AUC, but even reduces the AUC at higher lead times.

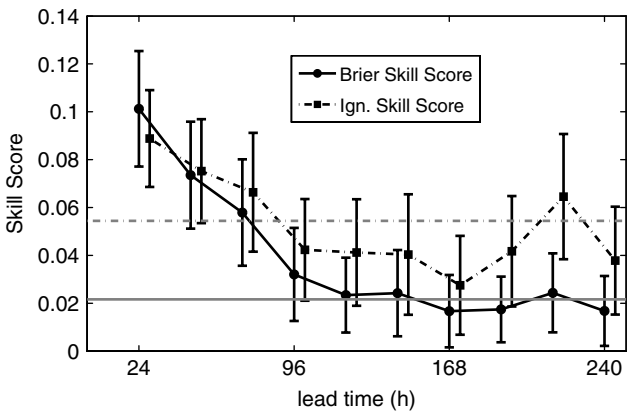

Figure 7. Skill scores for outlier prediction in the operational ensemble versus lead time. Larger values indicate better predictive skill. For short lead times, improvement is around 10% over base rate and saturates at less than 5% for lead times > 96 h. Skill scores significantly exceed zero for all lead times, as indicated by the 95% confidence bars. The Brier Skill Score agrees particularly well with the consistent case (horizontal grey lines) for lead times > 96 h.

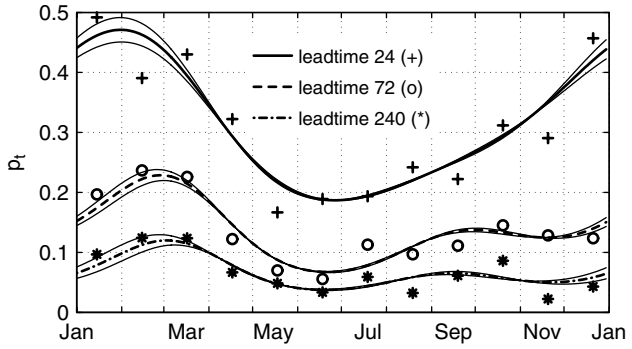

Figure 6. Estimated outlier probabilities of the logistic regression model for input combination 2. The envelopes indicate 95% confidence bands based on 10-fold cross-validation. Markers indicate relative outlier frequency per month in the corresponding ensemble. Outlier frequency is increased between December and May for all three cases. Model output and empirical monthly relative frequencies coincide.

Expressions for the variances and covariances in Eq. (15) can be found in DeLong *et al.* (1988). The constant base-rate forecast yields an AUC of 1/2. Input combination 1 given in Table I improves upon the base-rate forecast significantly for lead times smaller than 168 h. Combination 2 increases the AUC profoundly at lead time 168 h, and marginally for all other lead times. The inclusion of higher frequencies than 2ω was tested, but did not yield significant improvement. Adding the smallest and largest ensemble members as inputs (combination 3) adds to the AUC significantly for all lead times. The inclusion of the complete ensemble does not increase the AUC. The significant decrease of skill due to the inclusion of the whole ensemble at lead times later than 96 h is an indicator of lack of regularization by the L_2 penalty. A regularization technique that leads to more shrinkage such as the lasso (Hastie *et al.*, 2009) might yield better results in this case.

Finally, the optimal model coefficients are interesting from a diagnostic point of view. In Figure 5, we observed that the mere inclusion of the seasonally oscillating inputs increased the performance of the model. In Figure 6, we plot the outlier probability estimates of the model for input combination 2 (Table I) over the course of one year. The narrow confidence bands indicate good agreement among the cross-validation models. For comparison, observed outlier base rates for each month are shown. For the station considered, outlier frequency is increased, roughly, between December and May, for all lead times.

Note that, for a ROC analysis, it is not actually necessary that the forecasts are probabilities. It would suffice here to use a linear model instead of the logistic regression model. We have cross-checked our results by carrying out the same experiments using linear models with L_2-regularization. The results (not shown) were in excellent agreement with the results presented for logistic regression models.

4.2. *Skill scores*

The skill score S_{rel} is given by the improvement of the empirical score S over the empirical score S_{ref} of a reference forecast. In Eq. (6) an expression of a skill score is given using expectation values.

In practice, typically no closed form expression is available for the distribution of the variable of interest. The expectation value of a score therefore has to be approximated, for example by the corresponding empirical score over a test dataset, as given by Eq. (12). We employ the Brier Score as defined in Eq. (4) and Ignorance Score as defined in Eq. (5). We proceed by taking as the reference score the empirical score of the base-rate forecast, in which the unconditional base rate obtained from the training dataset is forecast at each time. Note that the base rate in the operational ensemble is in general larger than $2/(K+1)$.

Brier Skill Score and Ignorance Skill Score for outlier prediction by regularized logistic regression are shown in Figure 7. There is systematic dependence of skill on the lead time. The improvement over the base-rate forecast is larger for shorter lead times, independent of the scoring rule used to evaluate the predictions. For the Brier Skill Score, the agreement between consistent and operational ensemble is striking for lead times greater than 96 h. The agreement is not as pronounced for the Ignorance Skill Score. The apparent agreement of the Brier Skill Scores is

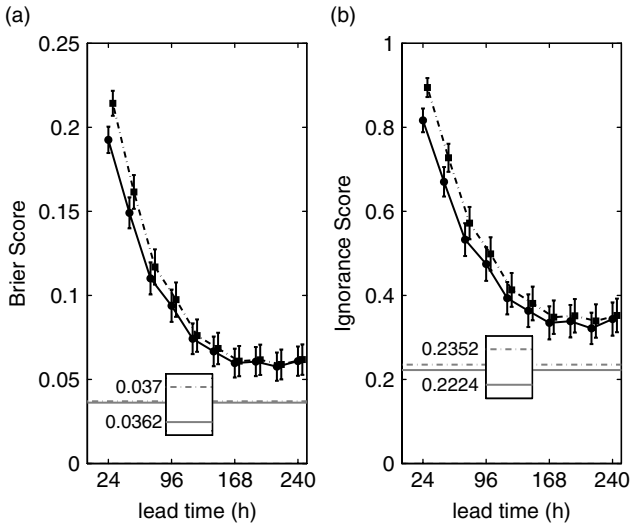

Figure 8. (a) Brier Scores for outlier prediction versus lead time. The solid line with 95% confidence bars indicates Brier Scores obtained by logistic regression in the operational ensemble. The dashed line with error bars indicates Brier Score of the base-rate prediction in the operational ensemble. The solid and dashed grey lines indicate the Brier Score of outlier prediction in the consistent ensemble based on the forecast distribution and based on the unconditional base rate, respectively. (b) is as (a), but for the Ignorance Score. Note that, unlike the Skill scores shown in Figure 7, these absolute scores indicate better prediction success by smaller values.

surprising, given that the predictive skill in the consistent case is based on knowledge of the values $e^{[1]}$, $e^{[K]}$, and the forecast distribution. In the operational ensemble though, no forecast distribution is available and the model uses the ensemble members as well as seasonally varying inputs in order to make the prediction. Figure 5 indicates that, without taking into account seasonality in the operational ensemble, predictive skill would be worse. Further, the agreement must not be over interpreted since even if the skill scores agreed the absolute scores could still be wrong by a common factor (see Eq. (6)).

The empirical means of the absolute Brier Score and Ignorance, as defined in Eqs (4) and (5), respectively, were evaluated for both the logistic regression estimates of the outlier probability as well as for the base-rate forecast in the operational ensemble. In Figure 8, it is shown that these absolute scores do not agree well with the scores expected in a consistent ensemble. For all lead times, scores obtained in the operational ensemble are worse than those of the consistent case, for both base-rate predictions and predictions based on the ensemble. This behaviour can be explained by the increased base rate of outliers in the operational ensemble. The Brier Score for predicting an event with base rate $\bar{\sigma}$ equals $\bar{\sigma}(1-\bar{\sigma})$ if the base rate is constantly issued as the forecast. This can be derived by analogy to Eq. (7). Hence, the Brier Score is monotonically increasing as a function of $\bar{\sigma}$ if $\bar{\sigma} < 0.5$, so that a base-rate forecast automatically yields better Brier Score with decreasing base rate. Furthermore, the scores of the ensemble-based prediction for the consistent case were obtained under the assumption that the forecast distribution is known, which is not the case in the operational ensemble. For these reasons, we should not expect to predict outliers in the operational ensemble with a better score than in the idealized consistent case. However, the *relative* increase of the ensemble-based prediction over the base-rate prediction is larger in the operational ensemble than it is in the consistent case, so that the *skill* scores turn out to be better in the operational ensemble. The dependence of the Brier and Ignorance Scores on the base rate also explains the lead-time dependence of scores in the operational ensemble. The base rate decreases with lead time (as indicated in Figure 6) and consequently the scores decrease, i.e. they become better.

We considered a number of additional inputs in the outlier prediction problem. These inputs included the full ensemble instead of only the outer members (cf. Figure 5), ensembles and high-resolution forecasts at different stations, forecast data at the same station of different variables (sea level pressure, wind speed), as well as forecast data from the same ensemble at different lead times. We did not find any more inputs other than season and outer ensemble members that significantly improved predictive skill in this particular ensemble. This is expected for a consistent ensemble but was not evident for an operational ensemble that is based on a complex model evolving in a large-dimensional phase space. Furthermore, we observed that using the outer ensemble members as predictors yielded the same skill as using only their difference (the ensemble range) and slightly better skill than using the inter-quartile range (IQR) or the standard deviation of the ensemble. The fact that the ensemble range is a better indicator for the occurrence of outliers than the IQR or the standard deviation is a consequence of the sensitivity of the range to fluctuations in the outer ensemble members. IQR and standard deviation are more robust to these fluctuations. For this reason, the ensemble range is more indicative of the probability concentrated outside the ensemble than the IQR or the standard deviation and thus serves as a better predictor for outliers.

It should be emphasized that we only provide a case-study using a single ensemble. In order to draw conclusions about the generality of our results, a thorough analysis of ensemble forecasts issued for different stations and variables by different institutions would be required. We briefly analyzed outlier prediction in the ECMWF temperature ensemble forecast issued for Heligoland (WMO 10015) and found positive skill scores that are comparable to those obtained for Dresden.

5. Conclusions

In this study we analyzed ensemble outliers, which we define as verifications lying outside the ensemble. In operational ensemble forecasts, such outliers are typically too frequent and thereby contribute to the lack of predictive skill. A better understanding of outliers is therefore of interest.

We show analytically that, in a consistent ensemble, better predictions of outliers can be issued than simply using the unconditional base rate. This result is based on the fact that, in a consistent ensemble, the ensemble members are independent draws from an underlying distribution; the performance of the outlier prediction turns out to be universal and dependent only on the number of ensemble members.

In the operational ensemble, outliers are likewise shown to be better predictable than just by their base rate. Even though the underlying probability distribution is not known, it is demonstrated that the ensemble contains enough information to issue skilful outlier predictions. The skill scores obtained for the operational ensemble are generally in good agreement with the theoretical results obtained for the consistent ensemble. This is surprising in so far as

the theoretical results assume that the forecaster knows the underlying distribution.

For the operational ensemble the outlier probabilities were estimated using regularized logistic regression. Firstly, the seasonal dependence of the occurrence of outliers justifies the use of ensemble interpretation methods that take into account the current season. Secondly, we apply techniques to distinguish important from unimportant predictors. We included the full ensemble instead of only the outer members, but found no significant improvement of predictive skill. In the same fashion, we included the same ensemble at different lead times, for different prognostic variables (such as mean-sea-level pressure and wind speed) and at neighbouring stations. No predictors beyond the ones presented in this study were found to improve predictive skill for the station we considered (WMO 10488). However, these results may vary for different stations.

This study provides arguments as to why the ordered ensemble members should be interpreted as order statistics, that is, as an ordered random sample from an underlying forecast distribution, rather than quantiles of that distribution. If the ensemble members were the quantiles, the mass of probability outside the outer members would be constant and equal to the unconditional base rate. If this were the case, the unconditional base rate would be the best possible probabilistic forecast for the occurrence of an outlier. Hence, there could be no increase in skill beyond the base-rate prediction, which is inconsistent with what we observe in our study. The analytical and empirical results presented in this study substantiate the interpretation and treatment of ensemble members as order statistics.

In summary, outlier events in ensemble forecasts deserve special attention as they represent unexpected events that are potentially harmful to an unprepared society. In future studies we plan to examine ensemble interpretation schemes with a focus on the potential for improvement with respect to unexpected events.

Acknowledgements

We are grateful to Renate Hagedorn and the European Centre for Medium-range Weather Forecasting for providing ensemble forecasts as well as station data for WMO 10488 and 10015. Questions and suggestions by two anonymous referees led to further improvements of the manuscript.

Appendix A. Distribution of outlier probability in a consistent ensemble

This appendix expands upon the discussion in section 2. The probability σ for the verification η being smaller than the first or larger than the Kth ensemble member, conditional on the ordered ensemble $E = (e^{[1]}, \ldots, e^{[K]})$ and the forecast cumulative distribution function $\Gamma(\cdot)$, is given by

$$\begin{aligned}\sigma &\equiv \Pr\left(\{\eta \le e^{[1]}\} \cup \{\eta > e^{[K]}\} \mid E, \Gamma\right)\\ &= \Gamma(e^{[1]}) + 1 - \Gamma(e^{[K]}).\end{aligned} \tag{A.1}$$

We call σ the outlier probability. Next we note that the probability integral transform of e_i, that is, the random variable $u_i \equiv \Gamma(e_i)$, is uniformly distributed over the interval $[0, 1]$ if the e_i are sampled randomly and independently from Γ (Mood *et al.*, 1974). Since Γ is a monotonically increasing function, the order of the samples is preserved between the e_i and u_i, that is $u^{[i]} = \Gamma(e^{[i]})$. We want to calculate the distribution of the quantity σ, for which we require the joint distribution of order statistics. For K randomly drawn uniformly distributed variables, the joint density of the ith and jth order statistic (with $i < j$) at values p and q, respectively, is given by (Balakrishnan and Rao, 1998)

$$\begin{aligned}f_{i,j}(p; q) =& \frac{K!}{(i-1)!\,(j-i-1)!\,(K-j)!}\\ &\times p^{i-1}(q-p)^{j-i-1}(1-q)^{K-j}.\end{aligned} \tag{A.2}$$

Setting $i = 1$ and $j = K$, the density $f_\sigma(s)$ of the outlier probability σ is given by

$$\begin{aligned}f_\sigma(s) &= p\left(u^{[1]} + 1 - u^{[K]} = s\right)\\ &= \int_0^s \mathrm{d}\tau\, f_{1,K}(\tau; \tau + 1 - s)\\ &= K(K-1)(1-s)^{K-2}s.\end{aligned} \tag{A.3}$$

Appendix B. Skill scores for outlier prediction in a consistent forecast ensemble

In order to calculate skill scores for the prediction of outliers in consistent forecast ensembles of size K, we first recall a property of the conditional expectation value (Mood *et al.*, 1974), namely

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}(X \mid Y)]. \tag{B.1}$$

Consider a scoring rule $s(p, y)$ that assigns a real number to a probabilistic forecast p for a binary event y. If the event happens, then $y = 1$ and otherwise $y = 0$. The scoring rule should reflect the quality of the single forecast p in predicting the event y. The score is given by the mathematical expectation $\mathbb{E}[s]$. It indicates the average performance of the forecast. By Eq. (B.1) the score can be written in terms of conditional expectation as $\mathbb{E}[s] = \mathbb{E}[\mathbb{E}(s \mid p)]$. Since the event y has only two possible outcomes (zero and one) which are mutually exclusive and collectively exhaustive, this can be further written as

$$\begin{aligned}\mathbb{E}[\mathbb{E}(s \mid p)] &= \mathbb{E}\big[s(p,1)\Pr(y = 1 \mid p)\\ &\qquad + s(p,0)\Pr(y = 0 \mid p)\big]\\ &= \mathbb{E}\left[s(p,1)p + s(p,0)(1-p)\right]\\ &= \int_0^1 \mathrm{d}\tau\left[s(\tau,1)\tau + s(\tau,0)(1-\tau)\right] f_p(\tau),\end{aligned} \tag{B.2}$$

where f_p is the density according to which p is distributed. The first equality in Eq. (B.2) follows because the forecast probability p is reliable. By conditioning the expectation, the expectation in Eq. (B.2) runs over p only and can be calculated by integrating over all τ, weighting the integrand by $f_p(\tau)$.

Commonly used scores for the assessment of probabilistic predictions are Brier Score and Ignorance Score, defined by Eqs (4) and (5), respectively. Solving the integral in Eq. (B.2) for the Ignorance Score yields

$$\begin{aligned}&\mathbb{E}\left[-p\log_2(p) - (1-p)\log_2(1-p)\right]\\ &\quad = \frac{1}{(K+1)\ln 2}\left\{2\sum_{j=3}^{K+1}\frac{1}{j} + \frac{(2K+1)(K-1)}{K(K+1)}\right\},\end{aligned} \tag{B.3}$$

where we have used the relations

$$\int_0^1 \mathrm{d}x\, x^{p-1}(1-x)^{q-1}\ln x = \frac{(p-1)!\,(q-1)!}{(p+q-1)!}\big\{\Psi(p)-\Psi(p+q)\big\}, \tag{B.4}$$

$$\Psi(n) = -\mathbf{C} + \sum_{j=1}^{n-1}\frac{1}{j}. \tag{B.5}$$

Here $\mathbf{C} = -0.577\ldots$ is the Euler–Mascheroni constant and we assume $p, q \in \mathbb{N}$ (Gradshteyn and Ryzhik, 1980). If we unconditionally predict the base rate $2/(K+1)$, the resulting Ignorance Score is

$$\mathbb{E}[s_{\mathrm{ref}}] = -\frac{2}{K+1}\log_2\frac{2}{K+1} - \frac{K-1}{K+1}\log_2\frac{K-1}{K+1}. \tag{B.6}$$

Skill scores are defined in Eq. (6). Combining Eqs (B.3) and (B.6), the Ignorance Skill Score is given by

$$\mathrm{IGN}_{\mathrm{rel}} = 1 + \frac{\left\{2\sum_{j=3}^{K+1}\frac{1}{j} + \frac{(2K+1)(K-1)}{K(K+1)}\right\}}{\left\{2\ln\frac{2}{K+1} + (K-1)\ln\frac{K-1}{K+1}\right\}}. \tag{B.7}$$

The ROC curve of the outlier probability σ is a plot of the hit rate $H(\zeta) = \Pr(\sigma > \zeta \mid y = 1)$ over the false alarm rate $F(\zeta) = \Pr(\sigma > \zeta \mid y = 0)$, where ζ is a varying threshold that parametrizes the ROC curve. We calculate the hit rate as follows.

$$\begin{aligned} H(\zeta) &= 1 - \int_0^\zeta \mathrm{d}\tau\, \Pr(\sigma = \tau \mid y = 1) \\ &= 1 - \frac{\int_0^\zeta \mathrm{d}\tau f_\sigma(\tau)\tau}{\Pr(y=1)} \\ &= \frac{1}{2}K(K+1)(1-\zeta)^{K-1}\zeta^2 \\ &\quad + (K+1)(1-\zeta)^K\zeta + (1-\zeta)^{K+1}. \end{aligned} \tag{B.8}$$

Similarly, we obtain the expression for the false alarm rate:

$$F(\zeta) = (K+1)(1-\zeta)^K\zeta + (1-\zeta)^{K+1}. \tag{B.9}$$

The area under the ROC curve (AUC) is given by

$$\int_0^1 \mathrm{d}F\, H = -\int_0^1 \mathrm{d}\zeta\, H\frac{\mathrm{d}F}{\mathrm{d}\zeta}.$$

Carrying out the calculations yields

$$\mathrm{AUC}(K) = \frac{11K^2 + 3K - 2}{16K^2 - 4}. \tag{B.10}$$

Appendix C. Outlier prediction Skill Scores and AUC in infinitely large ensembles

We consider the case $K \to \infty$. It is obvious that the Brier Skill Score, which is equal to $1/(K+2)$ as shown in Eq. (3), converges to zero as $K \to \infty$. For the Ignorance Skill Score, consider first

$$A_K \equiv 2\sum_{j=3}^{K+1}\frac{1}{j}. \tag{C.1}$$

Noting that

$$1 + \int_2^{K+2}\frac{\mathrm{d}x}{x-1} > \sum_{j=1}^{K+1}\frac{1}{j} > \int_1^{K+2}\frac{\mathrm{d}x}{x}, \tag{C.2}$$

we see that $A_K = 2\ln(K+1) + \mathcal{O}(1)$. Furthermore,

$$B_K \equiv \frac{(2K+1)(K-1)}{K(K+1)} \tag{C.3}$$

converges to 2 for $K \to \infty$. For the denominator of Eq. (B.7), let

$$C_K \equiv 2\ln\frac{2}{K+1} = -2\{\ln(K+1) + \mathcal{O}(1)\}, \tag{C.4}$$

and

$$\begin{aligned} D_K &\equiv (K-1)\ln\frac{K-1}{K+1} \\ &= \ln\left(1 - \frac{2}{K+1}\right)^{K+1} - \ln\left(1 - \frac{2}{K+1}\right)^2, \end{aligned} \tag{C.5}$$

which converges to -2 for $K \to \infty$. Combining these results, we get

$$\begin{aligned} \lim_{K\to\infty}\mathrm{IGN}_{\mathrm{rel}} &= 1 + \lim_{K\to\infty}\frac{A_K + B_K}{C_K + D_K} \\ &= 1 + \lim_{K\to\infty}\frac{2\{\ln(K+1) + \mathcal{O}(1)\}}{-2\{\ln(K+1) + \mathcal{O}(1)\}} \\ &= 1 - 1 = 0, \end{aligned} \tag{C.6}$$

demonstrating that the Ignorance Skill Score, too, converges to zero for $K \to \infty$.

Interestingly, the AUC does not converge to the trivial value of 1/2 for $K \to \infty$, but rather to a value of 11/16, as shown by Eq. (B.10).

References

Anderson J. 1996. A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate* **9**: 1518–1530.

Anderson J. 1997. The impact of dynamical constraints on the selection of initial conditions for ensemble predictions: Low-order perfect model results. *Mon. Weather Rev.* **125**: 2969–2983.

Balakrishnan N, Rao C. (eds.) 1998. *Handbook of Statistics 16: Order Statistics: Theory and Methods*. Elsevier: Amsterdam.

Barkmeijer J, Buizza R, Palmer TN. 1999. 3D-Var Hessian singular vectors and their potential use in the ECMWF Ensemble Prediction System. *Q. J. R. Meteorol. Soc.* **125**: 2333–2351.

Bowler N. 2006. Comparison of error breeding, singular vectors, random perturbations and ensemble Kalman filter perturbation strategies on a simple model. *Tellus A* **58**: 538–548.

Brier G. 1950. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **78**: 1–3.

Bröcker J. 2008. On reliability analysis of multi-categorical forecasts. *Nonlin. Proc. Geophys.* **15**: 661–673.

Bröcker J. 2010. Regularized logistic models for probabilistic forecasting and diagnostics. *Mon. Weather Rev.* **138**: 592–604.

Buizza R, Houtekamer P, Pellerin G, Toth Z, Zhu Y, Wei M. 2005. A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Weather Rev.* **133**: 1076–1097.

Buizza R, Palmer TN. 1998. Impact of ensemble size on ensemble prediction. *Mon. Weather Rev.* **126**: 2503–2518.

DeLong E, DeLong D, Clarke-Pearson D. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* **44**: 837–845.

Descamps L, Talagrand O. 2007. On some aspects of the definition of initial conditions for ensemble prediction. *Mon. Weather Rev.* **135**: 3260–3272.

Egan J. 1975. *Signal detection theory and ROC-analysis*. Academic Press.

Gneiting T, Raftery A. 2007. Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102**: 359–378.

Gradshteyn I, Ryzhik I. 1980. *Table of Integrals, Series and Products*. (corrected and enlarged ed.) Academic Press: New York, NY.

Hamill T, Colucci S. 1997. Verification of Eta–RSM short-range ensemble forecasts. *Mon. Weather Rev.* **125**: 1312–1327.

Hastie T, Tibshirani R, Friedman J. 2009. *The elements of statistical learning: Data mining, inference and prediction*. (2nd ed.) Springer: Berlin.

Kalnay E. 2003. *Atmospheric modeling, data assimilation, and predictability*. Cambridge University Press: Cambridge, UK.

Leith CE. 1974. Theoretical skill of Monte Carlo forecasts. *Mon. Weather Rev.* **102**: 409–418.

Mood A, Graybill F, Boes D. 1974. *Introduction to the theory of statistics*. (3rd ed.) McGraw–Hill: New York, NY.

Park M, Hastie T. 2008. Penalized logistic regression for detecting gene interactions. *Biostatistics* **9**: 30–50.

Roulston M, Smith L. 2002. Evaluating probabilistic forecasts using information theory. *Mon. Weather Rev.* **130**: 1653–1660.

Sivillo J, Ahlquist J, Toth Z. 1997. An ensemble forecasting primer. *Weather and Forecasting* **12**: 809–818.

Talagrand O, Vautard R, Strauss B. 1997. 'Evaluation of probabilistic prediction systems'. In *Proceedings of Workshop on Predictability*, ECMWF: Reading, UK.

Wilks DS. 2006. *Statistical methods in the atmospheric sciences*. (2nd ed.) Academic Press: New York, NY.

Ziehmann C. 2000. Comparison of a single-model EPS with a multi-model ensemble consisting of a few operational models. *Tellus A* **52**: 280–299.

