Adv. Sci. Res., 8, 53–57, 2012
www.adv-sci-res.net/8/53/2012/
doi:10.5194/asr-8-53-2012

Advances in Science & Research
Open Access Proceedings

On the predictability of outliers in ensemble forecasts

S. Siegert[1], **J. Bröcker**[1,*], **and H. Kantz**[1]

[1]Max Planck Institute for the Physics of Complex Systems, Dresden, Germany
[*]visiting research fellow: Centre for the Analysis of Time Series, London School of Economics, London, UK

Correspondence to: S. Siegert (siegert@pks.mpg.de)

Received: 27 September 2011 – Revised: 19 March 2012 – Accepted: 20 March 2012 – Published: 28 March 2012

Abstract. In numerical weather prediction, ensembles are used to retrieve probabilistic forecasts of future weather conditions. We consider events where the verification is smaller than the smallest, or larger than the largest ensemble member of a scalar ensemble forecast. These events are called outliers. In a statistically consistent K-member ensemble, outliers should occur with a base rate of $2/(K+1)$. In operational ensembles this base rate tends to be higher. We study the predictability of outlier events in terms of the Brier Skill Score and find that forecast probabilities can be calculated which are more skillful than the unconditional base rate. This is shown analytically for statistically consistent ensembles. Using logistic regression, forecast probabilities for outlier events in an operational ensemble are calculated. These probabilities exhibit positive skill which is quantitatively similar to the analytical results. Possible causes of these results as well as their consequences for ensemble interpretation are discussed.

1 Motivation

An *ensemble* is a collection of K individual forecasts for the same target. Here, we consider only forecasts of scalar variables, such as the temperature at a certain location. The individual ensemble members may differ in their initial conditions, in the model parameters, or they can even be produced by entirely different models. The objective of an ensemble forecast is to obtain an estimate of the flow-dependent forecast uncertainty, expressed, for example, in terms of a probability distribution function. The heterogeneities of the ensemble members reflect the lack of confidence in the initial state and the relevant physical processes. Under the special working hypothesis that all ensemble members are equally likely model scenarios, the final measurement – the *verification* – should behave like just another ensemble member in a statistical sense. An ensemble that satisfies this condition is called *statistically consistent* (Anderson, 1997).

Outliers in ensemble forecasts are events where the verification is smaller than the smallest or larger than the largest ensemble member (see Fig. 1). In this context, the term "outlier" was coined by Buizza and Palmer (1998). An ensemble of K-members defines $K+1$ intervals into which the verification can fall. If the ensemble is statistically consistent, none of these intervals should be preferred by the verification and each should occur with the same *base rate* (average relative frequency) of $1/(K+1)$. Two of the intervals correspond to outliers, hence the outlier base rate should be neither larger nor smaller than $2/(K+1)$ if the ensemble is consistent.

In operational forecast ensembles, however, outliers typically occur at a higher base rate. In Fig. 2 the outlier base rate in the ECMWF[1] temperature ensemble that was operational between 2001 and 2006 is shown. The location is Dresden, Germany (WMO10488). Throughout this study, we use ensemble data at the nearest grid point to this station and do not include the control run in the ensemble. For short-range forecasts outliers occur around 35 % of the time, for long lead times the outlier base rate saturates at around 10 %. If this 50-member ensemble were consistent, the outlier base rate should not exceed $2/51 \approx 3.9\,\%$. An increased frequency of outliers can have several causes such as conditional bias, lack of ensemble dispersion, model and discretization error, as well as observation noise (Hamill, 2001; Saetra et al., 2004).

Outliers are of interest because they can be interpreted as "unexpected" events if the range of possible scenarios is defined by the ensemble range. The outlier at lead time 1 day in Fig. 1 represents a particularly serious example of this interpretation. All of the 50 ensemble members predict

[1]European Centre for Medium-Range Weather Forecasts

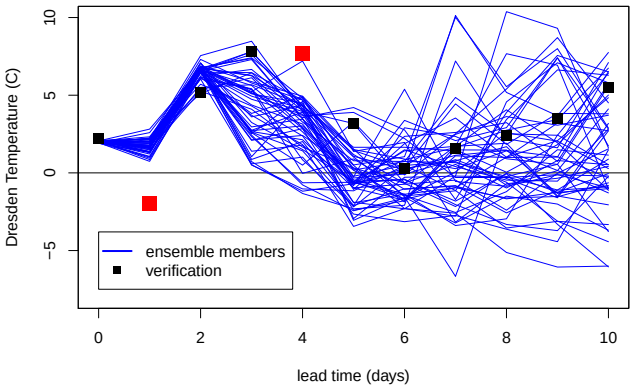

Figure 1. An ensemble temperature forecast issued by the ECMWF. Initialization date (lead time 0) is 22 December 2004, the location is WMO10488 (Dresden, Germany). The outliers at lead times 1 and 4 days are marked red.

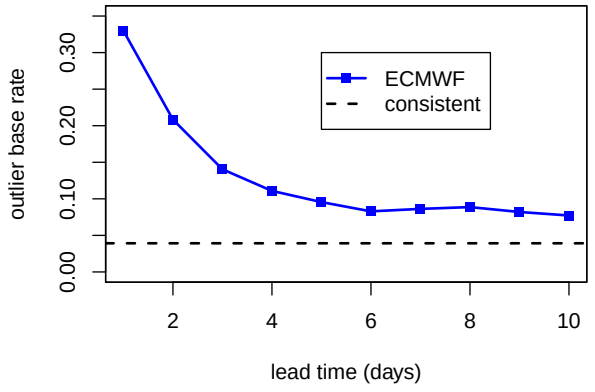

Figure 2. The markers indicate outlier base rate over lead time for the ECMWF temperature ensemble. The time period is 2001–2006 and the location is Dresden, Germany. The dashed line indicates the outlier base rate $2/51$ of a consistent 50-member ensemble.

temperature above zero degrees but the verification drops below zero. A forecast user might incur damage from this outlier not due to the extremeness of the event, but due to not being prepared for the ensuing weather conditions, such as freezing rain. Outliers (such as the ones in Fig. 1) are not necessarily extreme with respect to climatology, but they are extreme events with respect to the forecast distribution. Note that, in the present study, we do not consider the distance of the verification from the ensemble, but only whether it is inside or outside the ensemble range.

In summary, the increased frequency of occurrence of outliers on the one hand, and their interpretation as unexpected events on the other hand motivate us to consider their predictability. In particular we consider the following question: can probabilistic predictions be issued for outliers that are more skillful than their unconditional base rate of occurrence?

In Sect. 2 we review the Brier Skill Score which we use to assess the quality of probabilistic predictions. In Sect. 3 we provide theoretical arguments as to the predictability of outliers in a statistically consistent ensemble, thereby providing a benchmark result. This result is contrasted to prediction skill of outliers in an operational ensemble in Sect. 4. Section 5 concludes.

A similar study has recently been published in Siegert et al. (2011). The present study is different in its discussion of outlier predictability in the context of Brier Score decomposition, in the statistical prediction algorithm employed, a more careful discussion of the base rate effect, and in the details of the ensemble prediction system.

2 Evaluating probabilistic predictions by means of the Brier Skill Score

The *Brier Score* (abbr. BS, Brier, 1950) and *Brier Skill Score* (abbr. BSS, e.g. Wilks, 2006) provide means to evaluate and compare, respectively, the skill of probabilistic forecasts. For a single probabilistic prediction p of a binary event y the BS is given by

$$\mathrm{BS}(p,y) = (y-p)^2, \tag{1}$$

where the occurrence or non-occurrence of the event is coded by $y=1$ or $y=0$, respectively. The BS is negatively oriented and vanishes only for a perfect forecast. In the present study our goal is to compare the prediction skill of a forecast p to that of a base rate forecast $q=\bar{y}$. The BSS of p with respect to q is given by

$$\mathrm{BSS}(p,q) = 1 - \frac{\overline{\mathrm{BS}(p,y)}}{\overline{\mathrm{BS}(q,y)}}, \tag{2}$$

where the overbar indicates either the mathematical expectation value or an empirical average. In the present context, the numerical value of the BSS can be interpreted as the percentage of improvement of the forecast p over the base rate forecast.

3 Theoretical prediction skill in a consistent forecast ensemble

In this section a benchmark result as to the predictability of outliers in consistent forecast ensembles is derived. Two assumptions are made: the ensemble $\boldsymbol{e} = (e_1,\cdots,e_K)$ is statistically consistent and the cumulative distribution function $F(\cdot)$, from which ensemble members and the verification η are independently drawn, is known to the forecaster. If F and $\boldsymbol{e}$ are known, the "true" outlier probability σ can be calculated by

$$\sigma := \mathbb{P}(\eta \notin [e_{[1]},e_{[K]}] \mid F,\boldsymbol{e}) = F(e_{[1]}) + 1 - F(e_{[K]}), \tag{3}$$

where $e_{[i]}$ denotes the i-th order statistic, that is, the value of the i-th largest ensemble member. Obviously, σ is not a constant but a fluctuating random quantity. In Siegert et

al. (2011) we show that the probability distribution function (pdf) of σ, evaluated at the value x, is given by

$$p_\sigma(x) = K(K-1)(1-x)^{K-2}x, \tag{4}$$

and thus independent of F. The universality of the pdf of σ is a result of the fact that $F(x)$ is uniformly distributed on the unit interval if x is a random sample drawn from F (Mood et al., 1974). Note that under the distribution given by Eq. (4), the expectation value of σ (the base rate) is correctly given by $2/(K+1)$.

For the calculation of the BSS the expectation value of the BS of σ, i.e. $\mathbb{E}(y-\sigma)^2$, as well as the expectation value of the BS of the base rate forecast, i.e. $\mathbb{E}(y-\frac{2}{K+1})^2$ are required. The calculation of the BS of σ involves conditional expectations (Mood et al., 1974) and the equality $\mathbb{P}(y=1\,|\,\sigma)=\sigma$. One can show that $\mathbb{E}(y-\sigma)^2 = \mathbb{E}\sigma(1-\sigma)$, which can then be explicitly calculated using Eq. (4). Combining the two resulting Brier Scores using Eq. (2) yields for the BSS of outlier prediction

$$\mathrm{BSS}\left(\sigma, \frac{2}{K+1}\right) = \frac{1}{K+2}, \tag{5}$$

under the assumptions of a consistent ensemble and a known forecast distribution. We conclude that under this idealized scenario, outliers are indeed better predictable than merely by their unconditional base rate. The result is universal since it is independent of the forecast distribution. The improvement over the base rate forecast is the higher, the smaller the ensemble is and vanishes as $K \to \infty$. The positive skill is obtained by predicting the true (fluctuating) outlier probability instead of its mean. In the 50-member ECMWF-ensemble, for example, we would get a BSS of close to 2 % improvement over the base rate forecast under the idealized scenario.

As it was shown in Murphy (1973), the Brier Score can be decomposed into a sum of three terms, namely an *uncertainty* term, a property only of the process to be predicted, a *reliability* term which quantifies the deviance of the forecast from reliability, and a *resolution* term which quantifies how much the individual predicted probabilities differ from each other. In fact, this decomposition is possible for every proper score (Bröcker, 2009). In the outlier prediction problem, the σ-forecast obtains its improvement in skill from a non-vanishing resolution term – the uncertainty terms of the σ-forecast and the base rate forecast are equal and their reliability terms vanish. The pdf of σ given by Eq. (4) is a Beta distribution whose variance decreases as the ensemble size K increases. In fact, for any reliable forecast (such as σ here), the resolution term of the BS is given by the variance of the forecast probabilities. Thus, the vanishing skill of σ resulting from a decreasing resolution can equivalently be explained by a decreasing variance of σ which leads to the σ-forecast becoming more and more similar to the base rate forecast. This example provides an intuitive illustration of the concept of resolution in probabilistic forecasting.

4 Outlier prediction skill in an operational ensemble

In an operational setting the assumptions that lead to the universal result given by Eq. (5) are not satisfied. Neither is the ensemble consistent (see Fig. 2), nor is the true distribution from which the ensemble is drawn known to the forecaster. Notwithstanding that, outlier probabilities that are conditional on the present state of the ensemble can still be estimated by means of statistical learning.

We trained a *logistic regression* model with Lasso regularization (Tibshirani, 1996) that takes as its input vector $\boldsymbol{x}_t$ at time instance t a constant intercept, both the smallest and the largest ensemble member and also seasonal varying inputs with frequency $\phi = 2\pi/(365\,\mathrm{days})$:

$$\boldsymbol{x}_t = (1, \sin(\phi t), \cos(\phi t), e^t_{[1]}, e^t_{[K]}). \tag{6}$$

Logistic regression fits a linear superposition of the inputs to the log-odds ratio of the outlier probability p_t, that is

$$\boldsymbol{x}_t\boldsymbol{\beta} = \log\left(\frac{p_t}{1-p_t}\right) \tag{7}$$

where $\boldsymbol{\beta}$ is the vector of coefficients. The coefficients are obtained using cyclical coordinate descent to maximize the log-likelihood of the data given the coefficients (Friedman et al., 2010). The Lasso penalty leads to shrinkage of unimportant inputs, thus avoiding overfitting. We apply 10-fold cross-validation to estimate out-of-sample performance. Our data set comprises 6 yr of daily temperature ensemble forecasts ($K = 50$) for Dresden (WMO10488) issued by the ECMWF ensemble between 2001 and 2006.

At each instance we issue the Lasso-estimated outlier probability p_t. Simultaneously we issue the constant base rate forecast, where we use as the base rate the outlier frequency known from the training data. The Brier Scores of these predictions are compared using the BSS.

The BSS over lead time is shown in Fig. 3. At all lead times, BSS is significantly greater than zero, indicating significant predictive skill of the Lasso-estimated outlier probabilities. At short lead times (for which a lot of outliers occur), the BSS is about 10 % improvement over the base rate forecast. This is a large improvement compared to the analytical result of about 2 %. The ensemble at short lead times has an increased outlier base rate and thus, qualitatively, behaves like an ensemble with fewer than 50 members. Since, according to Eq. (5), the BSS is larger in smaller ensembles, the observed skill in the short lead time ensembles is larger than the analytical BSS using $K = 50$. At higher lead times, the BSS saturates at a small, but significantly positive, value which is close to the analytical result for the consistent ensemble, despite the operational ensemble not being consistent. This observation does not change if we apply statistical downscaling of the ensemble using the four neighboring grid points, instead of using ensemble data only at the nearest grid point. Our downscaling corrects for seasonal bias, which reduces the outlier base rate shown in Fig. 2 by around 3 % at

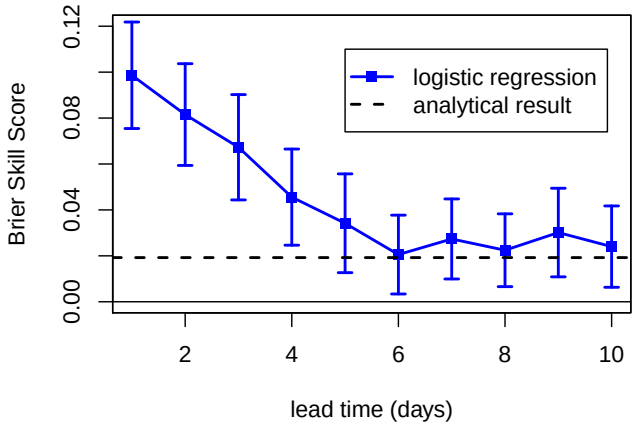

Figure 3. Brier Skill Score over lead time of estimated outlier probability compared to base rate forecast. Error bars indicate 95 % confidence intervals. For reference, the analytical result obtained in Sect. 3 using $K = 50$ is shown.

all lead times. But the behavior of Brier Skill Score shown in Fig. 3 remains the same. Further, the BSS is invariant if we use ensemble and verification data at a different station (Heligoland, WMO10015) and if we include the control as an additional ensemble member. Note, however, that the results presented here might not be representative of the ECMWF ensemble which is currently operational.

We have tested further potential predictors for the logistic regression estimates. Examples include the full ensemble $\boldsymbol{e}^t$, higher frequencies of the seasonal inputs, ensembles at neighboring stations, ensembles at smaller and larger lead times and ensembles of different variables at the same station. None of these could significantly improve the predictive skill over that of the predictors given in Eq. (6).

5 Summary and conclusions

We have considered the predictability of outliers in ensemble forecasts. The main result is that outliers are predictable better than by their unconditional base rate. This was shown in terms of the BSS both for a hypothetical ensemble which is perfectly statistically consistent, and for an operational temperature forecast ensemble. For both ensembles, the predictive skill is obtained from the fact that the outlier probability is conditional on the current state of the ensemble. In the operational ensemble, additional skill is obtained from seasonality.

A more detailed analysis shows that outliers become more likely if the range of the forecast ensemble is low. In fact, using the range as a predictor instead of $e^t_{[1]}$ and $e^t_{[K]}$ does not alter the BSS significantly. The fact that a small range leads to a higher outlier probability seems trivial but at the same time it contradicts to some extent the common view that a very narrow ensemble is an indicator of low forecast uncertainty. Sampling effects can lead to an increase or decrease of the outlier probability compared to the base rate.

The empirical results of Sect. 4 suggest that the members of the operational ensemble should be interpreted as order statistics drawn from a distribution and not as quantiles of a distribution. If ensemble members were quantiles, the mass of probability concentrated between the ensemble members, and outside the ensemble, would be a constant. The outlier probability conditional on the ensemble would then be equal to the unconditional base rate. However, by allowing this probability to change as a function of certain characteristics of the ensemble we obtain significant forecast skill.

A combination of an increased base rate, seasonality, and the performance of an empirical statistical learning algorithm leads to approximately the same forecast skill in the operational ensemble that we expect if the ensemble were consistent and if we knew the distribution from which the members were drawn. Neither of the latter assumptions is satisfied in operational ensembles. Based on these differences, we should not overinterpret the apparent similarity in predictive skill. The main result is that predictive skill of outlier events in the operational ensemble is small, yet significant. However, the analytical benchmark skill score made the observed skill more interpretable, which motivates us to further address predictability problems in that way.

Acknowledgements. We are grateful to Renate Hagedorn and the ECMWF for providing ensemble and station data for Dresden and Heligoland. We thank Christopher Ferro and an anonymous referee for helpful comments on an earlier draft of this article.

The service charges for this open access publication have been covered by the Max Planck Society.

Edited by: H. Böttger
Reviewed by: A. Persson and C. Ferro

References

Anderson, J.: 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, 1997.

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

Bröcker, J.: Reliability, sufficiency, and the decomposition of proper scores, Q. J. Roy. Meteorol. Soc., 135, 1512–1519, 2009.

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

Friedman J., Hastie, T., and Tibshirani, R.: Regularization Paths for Generalized Linear Models via Coordinate Descent, J. Stat. Softw., 33, 1–22, 2010.

Hamill, T. M.: Interpretation of rank histograms for verifying ensemble forecasts, Mon. Weather Rev., 129, 550–560, 2001.

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

Murphy, A. H.: A new vector partition of the probability score, J. Appl. Meteorol., 12, 595–600, 1973.

Saetra, Ø., Hersbach, H., Bidlot, J. R., and Richardson, D. S.: Effects of observation errors on the statistics for ensemble spread and reliability, Mon. Weather Rev., 132, 1487–1501, 2004.
Siegert, S., Bröcker, J., and Kantz, H.: Predicting outliers in ensemble forecasts, Q. J. Roy. Meteorol. Soc., 137, 1887–1897, 2011.
Tibshirani, R.: Regression shrinkage and selection via the Lasso, J. Roy. Stat. Soc. B, 58, 267–288, 1996.
Wilks, D. S.: Statistical methods in the atmospheric sciences (2nd Edn.), Academic Press, New York, NY, 2006.