METEOROLOGICAL APPLICATIONS
Meteorol. Appl. **15**: 155–162 (2008)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/met.50

How much does simplification of probability forecasts reduce forecast quality?

F. J. Doblas-Reyes,[a]* C. A. S. Coelho[b] and D. B. Stephenson[c]

[a] *European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading RG2 9AX, UK*
[b] *Centro de Previsão de Tempo e Estudos Climáticos, Rodovia Presidente Dutra, Km 40, SP-RJ 12630-000, Cachoeira Paulista, Brazil*
[c] *School of Engineering, Computing and Mathematics, University of Exeter, Harrison Building, North Park Road, Exeter EX4 4QF, UK*

ABSTRACT: Probability forecasts from an ensemble are often discretized into a small set of categories before being distributed to the users. This study investigates how such simplification can affect the forecast quality of probabilistic predictions as measured by the Brier score (BS). An example from the European Centre for Medium-Range Weather Forecasts (ECMWF) operational seasonal ensemble forecast system is used to show that the simplification of the forecast probabilities reduces the Brier skill score (BSS) by as much as 57% with respect to the skill score obtained with the full set of probabilities issued from the ensemble. This is more obvious for a small number of probability categories and is mainly due to a decrease in forecast resolution of up to 36%. The impact of the simplification as a function of the ensemble size is also discussed. The results suggest that forecast quality should be made available for the set of probabilities that the forecast user has access to as well as for the complete set of probabilities issued by the ensemble forecasting system. Copyright © 2008 Royal Meteorological Society

KEY WORDS forecast quality; probability forecasts; ensemble forecasting

Received 21 September 2007; Revised 18 December 2007; Accepted 2 January 2008

1. Introduction

One of the scores that has been most widely used to measure the quality of weather and climate probabilistic forecast systems is the Brier score henceforth (BS) [Brier, 1950]. Using a decomposition into three components (Murphy, 1986), this score measures different quality aspects of probability forecasts: reliability (bias of conditional means), resolution (variance of conditional means), and observational uncertainty. The decomposition can be achieved by using the probability forecasts from either the complete set of probability values issued by the forecast system or a simpler, smaller set of probability categories. Each method has its advantages and it is traditionally the forecast producer who decides how to estimate the BS and its components.

The reliability and resolution terms of the BS decomposition have been widely used in the context of operational ensemble forecast systems (e.g. Atger, 2004). Traditionally, reliability and resolution have been made available based on estimates that do not discriminate between the different probabilistic forecast products offered to the users. For instance, for simplicity the most recent seasonal forecasting system at ECMWF offers maps of probabilities for specific events (e.g. temperature above the upper tercile) in seven categories of probability: 0–0.1, 0.1–0.2, 0.2–0.4, 0.4–0.5, 0.5–0.6, 0.6–0.7 and 0.7–1. This is a number smaller than the set of 42 probability values that can be produced from the 41-ensemble members provided by the ECMWF coupled atmosphere-ocean ensemble forecast system with simple, frequentist, non-parametric methods. There are many reasons for the simplification, such as to smooth out the probabilities issued or to avoid the sparseness of some probability categories. Probability forecasts are also simplified in order to issue severe weather warnings, as happens at the UK Met Office. Early warnings are issued when the probability of an extreme event exceeds 0.6. Such warnings are formulated using a crude categorization of the forecast probabilities. The forecast quality of these examples should be estimated for the set of simplified probability categories rather than for the full set of probabilities without simplification.

Following Stephenson *et al.* (2008), the reliability and resolution terms calculated from the set of seven probability categories of the ECMWF forecasts is expected to be different to the reliability and resolution components calculated using the full set of 42 probability categories. More importantly, the BS computed from the simplified set of forecast probabilities should be larger than the BS computed for the full set, i.e. the simplification may lead to a reduction in forecast quality. This study shows how large the reduction in forecast quality is expected to be for a typical operational seasonal forecasting system, although the effect of the simplification of the probability forecasts is more general and affects predictions issued

* Correspondence to: F. J. Doblas-Reyes, European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading RG2 9AX, UK. E-mail: f.doblas-reyes@ecmwf.int

with all sorts of probabilistic forecast systems. Section 2 describes the ensemble forecasts used for the illustration, while Section 3 shows some examples of the effect of the simplification on the BS and its decomposition. Section 4 contains a summary of the results and some recommendations.

2. Data

This study uses operational seasonal ensemble forecasts from the most recent system developed at ECMWF, known as System 3 (Anderson *et al.*, 2007). The system is based on a dynamical coupled model with IFS cycle 31R1 and HOPE-E as atmospheric and ocean components, respectively. The forecasts are initialized once *per* month with the ECMWF operational analyses for the atmospheric and soil variables and with the System 3 ocean analyses (Balmaseda *et al.*, 2007) for the ocean variables. The 41-member ensemble is generated using wind-stress and sea surface temperature (SST) perturbations, along with atmospheric singular vectors. The forecasts have a length of 7 months. To estimate the forecast quality and to allow for calibration of the forecasts, a set of re-forecasts, or hindcasts, have been carried out for the period 1981–2005. The hindcasts are created in an identical way to the real-time operational forecasts, but with 11-member ensembles.

ERA40 (Uppala *et al.*, 2005) and ECMWF operational analyses have been used as reference dataset for the atmospheric fields. In spite of the limitations of reanalysis data being used as surrogates for observations, they ensure homogeneity and global coverage of the fields.

3. The Brier score and its decomposition in an operational context

Ensemble forecasts have been widely used to issue probability forecasts (e.g. Richardson, 2001), although they are not the only method available for this purpose (Stephenson *et al.*, 2005). In the case of a dichotomous event, given an ensemble of simulations, a simple way of obtaining a probability forecast consists of computing the fraction of ensemble members for which the value of a given variable exceeds a given threshold. More sophisticated methods to obtain the full forecast probability distribution function from the ensemble have been proposed (e.g. Roulston and Smith, 2003; Stephenson *et al.*, 2005), but given the limited sample size of seasonal forecasts, a simple, frequentist, non-parametric approach has been used. For long-range forecasts such as those used in this study, the thresholds are usually defined in terms of percentiles of the climatological distribution. In the following, the threshold that defines the forecast event is chosen separately for the verification dataset and the set of forecasts. This takes account of the known and unavoidable systematic errors of present-day coupled models. This method is similar to the 'bias-corrected relative frequency' used by Hamill and Whitaker (2006).

Computing the forecast probabilities as a fraction of the ensemble members satisfying a threshold-based criterion implies that the maximum set of probabilities issued is determined by the ensemble size plus one. For ECMWF System 3 operational seasonal forecasts, 41-member ensemble forecasts could be used to create a maximum of 42 different forecast probabilities: 0, 1/41, 2/41, ... , 40/41 and 1. However, it is common to simplify forecasts using a smaller number of probability categories. For instance, System 3 operational seasonal forecast maps for the event 'temperature above the upper tercile' are offered with seven probability categories. These probability categories are obtained by categorizing the full set of 42 different probability values. Other examples of the simplification of probability forecasts when preparing products for users can be found at the International Research Institute for Climate and Society (Barnston *et al.*, 2003). Stephenson *et al.* (2008) show that both the value of the BS and its decomposition depend on the set of probabilities used in the calculation. This suggests that the scores should be estimated for each set of probability categories provided to the users. In other words, each probability forecast product with a different set of probability categories, as the ones described above, requires an individual estimate of the BS along with the reliability and resolution terms, which should be available to the user with the forecast.

Attribute diagrams (Hsu and Murphy, 1986) allow the visualization of the reliability, resolution and sharpness of a set of forecasts for a specific dichotomous event. They are used here to illustrate the implications of providing operational products with different sets of probability categories. Figure 1 shows the attribute diagrams for 1-month lead seasonal forecasts of 2-m temperature over the tropics started on the 1st of May. The diagrams illustrate the conditional relative frequency of occurrence of the event as a function of the forecast probability. Results for each probability interval obtained from all the probability forecasts falling inside the interval are represented by a solid circle whose area is proportional to the sample size, i.e. the number of probability forecasts in the interval over all the years and grid points of the region. The set of categories is (0,1/3), (1/3,2/3) and (2/3,1) for the diagram with three categories and so on until (0,1/12), (1/12,2/12), ..., (11/12,1) for twelve categories. As categories are intervals of probability, a representative probability for the category should be defined. This is also the point on the abscissa where the result for a given category is drawn in the attribute diagram. The representative forecast probability for the category used in Figure 1 is a weighted mean of the individual forecast probabilities included in the category. This is different from the defining of the representative forecast probability as the centre of the probability interval of the category, which is what most centres actually do in their operational forecasts. Nevertheless, we have chosen the former option because it illustrates the effect of the simplification in a clearer way. The

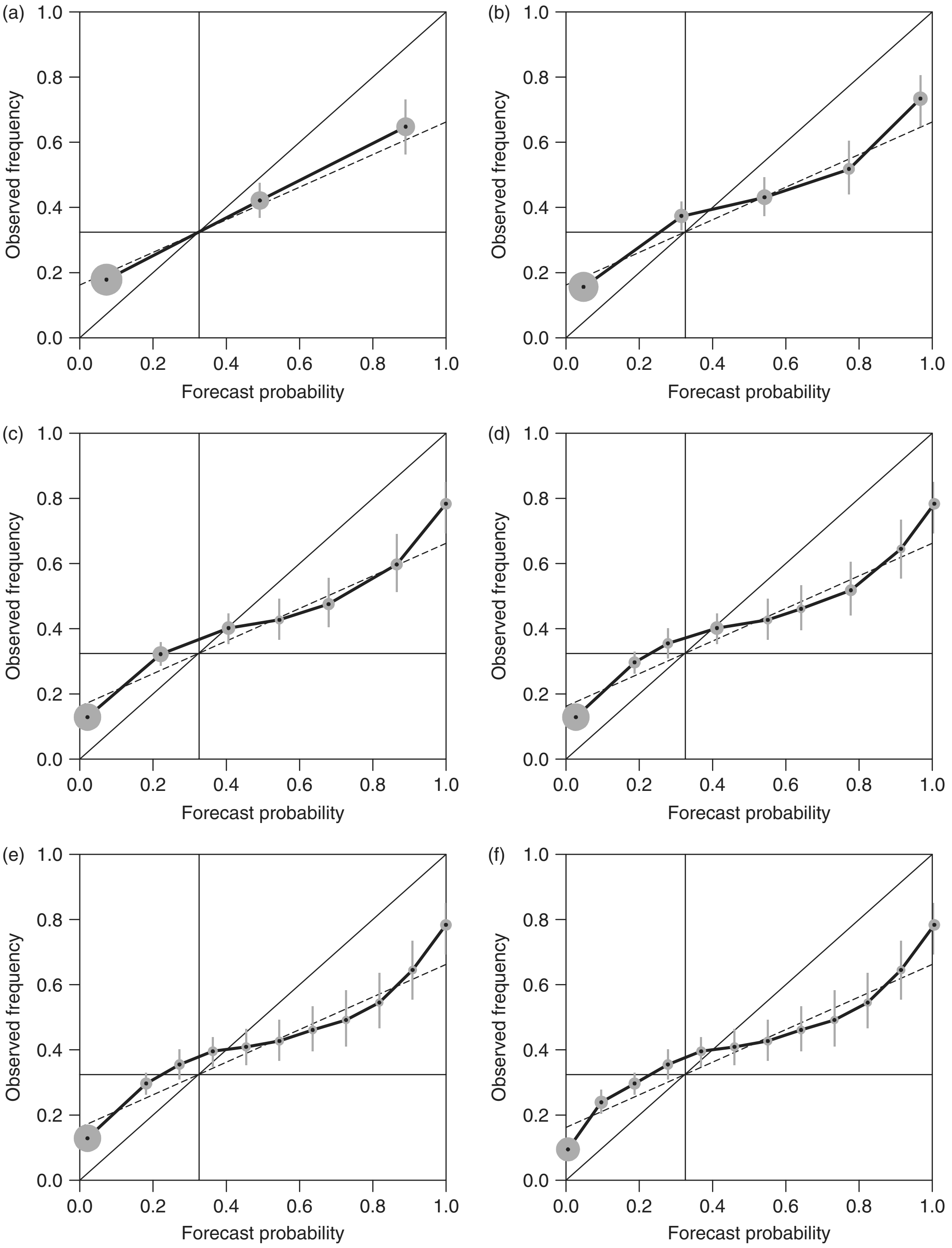

Figure 1. Attribute diagrams of one-month lead seasonal forecasts of summer temperature above the upper tercile over the tropics (20°N–20°S) started on the 1st of May during the period 1981–2005 for a) 3, b) 5, c) 7, d) 9, e) 11 and f) 12 probability categories. Each forecast probability category is represented by a solid circle whose area is proportional to the sample size of the category. The vertical solid line represents the average forecast probability, while the horizontal solid line is for the observed climatological frequency of the event over the period 1981–2005. The dashed line separates skilful from unskilful regions in the diagram: points with forecast probabilities smaller (larger) than the climatological frequency that fall below (above) this line, contribute to a positive Brier skill score (BSS); otherwise they contribute negatively to the BSS. Grey vertical bars over the dots indicate the 90% confidence intervals of the estimated observed frequency based on a 1,000 bootstrap resampling procedure.

consequences of this choice are discussed in more detail in Section 4.

In the idealized case of infinite sample and ensemble sizes, the diagonal line represents the results for a set of forecasts with perfect reliability. This occurs if, from the cases where an event is forecast with probability p, the event occurs on a fraction p of occasions. In practice, when an ensemble forecast system predicts some event with probability p, the event occurs in reality in a fraction q of times. If p is different from q, the probability

DOI: 10.1002/met

forecasts obtained from the ensemble are not reliable. This situation will appear in the diagram as a point away from the diagonal. If the corresponding curve is shallower than the diagonal the forecast system is said to be overconfident, while if it is steeper the system will be underconfident. The sum of the horizontal square distance of all the points to the diagonal (weighted by the sample size of each probability category) is an estimate of the lack of reliability of the system as measured by the BS. In the same way, the sum of the vertical distance of the points to the horizontal line corresponding to the climatological frequency of the event measures the forecast resolution, i.e. the ability of the system to issue reliable forecasts different from the naive climatological probability (the horizontal line at 1/3 observed frequency in the examples). This means that if the reliability curve were to be horizontal, the frequency of occurrence would not depend on the forecast probabilities and the system would have zero resolution (and no skill over a climatological forecast). The dashed line in the diagram separates skillful from unskillful regions in the diagram: only points with forecast probabilities smaller (larger) than the climatological frequency which fall below (above) this line, contribute to a positive Brier skill score (BSS) with respect to a climatological forecast. The BSS is a measure of the relative benefit of the forecasts with respect to using the naïve climatological probabilities and is defined as $\mathrm{BSS} = 1 - \mathrm{BS}/\mathrm{BS}_c$, where BS_c is the BS of the climatological forecast, the one that always issues as forecast probability the historical frequency of the event. Finally, the sharpness is an estimate of the variance of the forecast probabilities. A visual estimate of the sharpness can be obtained from the diagram by considering how far and large the dots are away from the average forecast probability (the vertical line at 1/3 forecast probability in the example).

The diagrams in Figure 1 show important changes for forecasts with less than five probability categories. In those cases, the resolution is limited by the closeness of the dots to the climatological frequency. When the number of categories is larger than five, the number of probability categories with observed frequency away from the climatological frequency increases, which also increases the sharpness (not shown). As a consequence, increasing the number of categories increases the forecast resolution. However, large sharpness is an extremely undesirable feature in the case of overconfident curves, such as the ones shown here. A level of 90% confidence intervals (grey bars) for the attribute diagram has been computed using a bootstrap method, where the original 11-member ensemble hindcasts (and the corresponding verification data) were re-sampled with replacement 1000 times (Nicholls, 2001; Lanzante, 2005; Jolliffe, 2007). The confidence intervals are away from the climatological frequency line for most of the probability categories, an indication that the forecasts have statistically significant resolution. The bars are in most cases away from the diagonal, suggesting that the probability forecasts issued with this simple method from the System 3 ensembles are unreliable and should be subject to some form of calibration before use.

The increase of resolution with the number of probability categories is further illustrated in Figure 2(a). The resolution (grey line) increases rapidly up to six categories, beyond which it saturates until it reaches 11 categories, where it shows a gap compared to the resolution of the

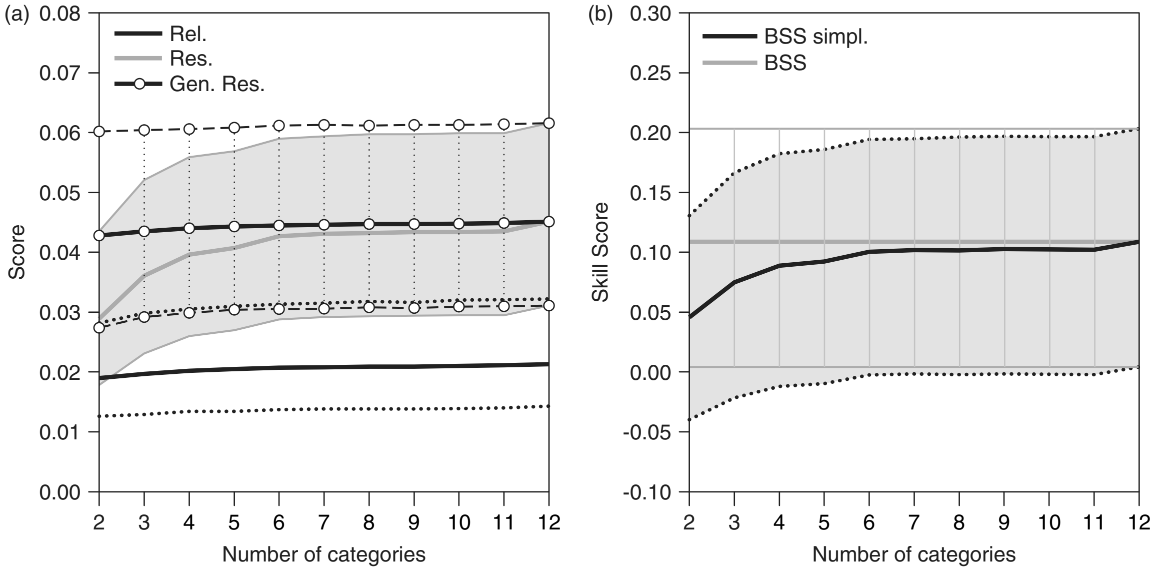

Figure 2. (a) Reliability (black), resolution (grey) and generalized resolution (black with circles) for one-month lead seasonal forecasts of summer temperature above the upper tercile over the tropics (20°N–20°S) started on the 1st of May during the period 1981–2005. 90% confidence intervals are displayed with black dotted, grey solid and dashed with circles lines for the reliability, resolution and generalized resolution, respectively. For clarity purposes the interval for the resolution has been shaded. (b) Brier skill score (BSS) with respect to climatology from the full set of forecast probabilities (grey solid line) and from simplified probabilities (black solid line) as a function of the number of categories for the same set of forecasts. 90% confidence intervals are displayed with grey solid and black dotted lines for the BSS from the full set and from the simplified probabilities, respectively. For clarity purposes the interval for the BSS from the simplified probabilities has been shaded.

full set of probabilities. This means that forecasts issued with less than six probability categories will suffer from a reduced resolution, regardless of the ensemble size of the forecast system. The reliability term also increases with the number of categories, although much less than the resolution. The reader is reminded that the reliability term is negatively oriented, so that an increase implies a worsening of the forecast reliability. A level of 90% confidence intervals using the same re-sampling method described above is also shown. Besides that, the reliability and the resolution are affected by the simplification. Figure 2(b) shows the BSS obtained using simplified probabilities (black solid line), which can be compared to the BSS obtained without simplification (grey solid line). The skill scores show very large confidence intervals, illustrating the large uncertainty in the forecast quality estimates that both forecasters and users have to deal with. For all numbers of categories lower than 12 (which is the maximum possible number of categories for an ensemble system with 12 ensemble members), the BSS for simplified probabilities shows lower skill than the BSS obtained without simplification, especially for less than six categories. Note that the BSS obtained without simplification is statistically significantly different from zero at the 10% level, while the BSS obtained using categorized probabilities is statistically significantly different from zero only if seven or more categories are used. The reduction is mostly due to the reduced resolution of the simplified probability forecasts.

Stephenson *et al*. (2008) found that the three-term decomposition of the BS will equal the BS obtained without simplification only if all the possible values of forecast probability are used. When the set of probability values is simplified, the BSS decreases as shown above. They found that two additional terms, which account for the variance and covariance of the probabilities included in each category [see Equation (4) in Stephenson *et al*. (2008)], are required to ensure that the BSS obtained with and without simplification are the same. The terms are added to the resolution component to define a generalized resolution that is less sensitive to the choice of the number of categories. Figure 2(a) shows the generalized resolution as a function of the number of categories. The generalized resolution has a much smaller dependence on the number of categories than the resolution. As the number of categories increases, the generalized resolution also increases to compensate the increase in the reliability term to keep the BS independent of the number of categories. The difference between the resolution and generalized resolution terms gives an idea of how much accuracy decreases by simplifying the full set of forecast probabilities.

The previous results have been obtained using the 11-member ensemble hindcasts of System 3. However, the ECMWF operational seasonal forecasts are formulated with a 41-member ensemble. Hence, the large difference in ensemble size between hindcasts and forecasts might be an additional source of error for the estimated BSS and its decomposition. To assess the impact of the ensemble size, the BSS has been computed for probability forecasts obtained using subsets of ensemble members. Figure 3 shows the impact of the ensemble size in the BSS of the simplified forecasts as a function of the number of categories. Note that for a given ensemble size, the maximum number of categories is the ensemble size plus one. This is the reason for some lines to be shorter than others.

Forecasts with an ensemble size smaller than nine show large variations of the BSS with the number of members. This is found for both versions of the BSS, with and

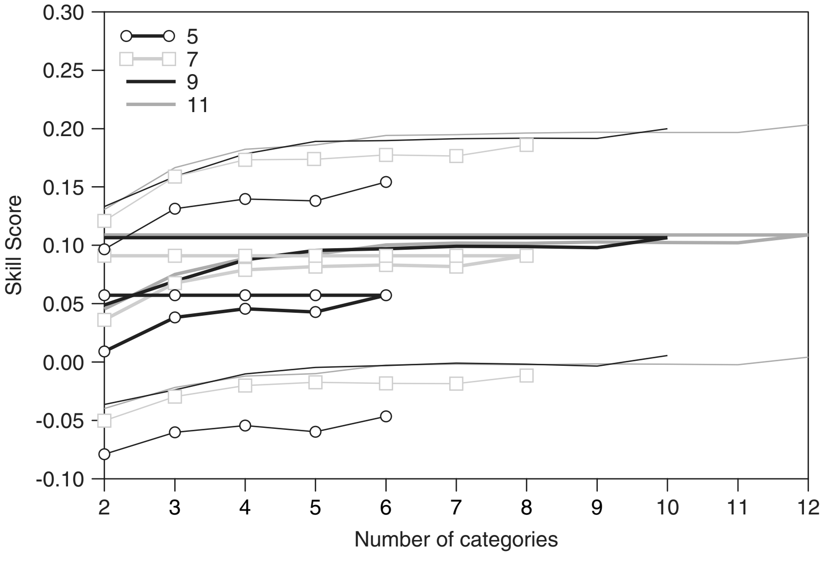

Figure 3. Brier skill score (BSS) with respect to climatology for the simplified probabilities (thick lines) and for the full set of probabilities (i.e., without simplification; thick horizontal lines) as a function of the number of probability categories for different ensemble sizes: 5 member, black line with circles; 7 members, grey line with squares; 9 members, black solid line; 11 members, grey solid line). The results are for one-month lead seasonal forecasts of summer temperature above the upper tercile over the tropics (20°N–20°S) started on the 1st of May during the period 1981–2005. 90% confidence intervals for the BSS of the simplified probabilities are displayed with thinner lines of the corresponding style.

DOI: 10.1002/met

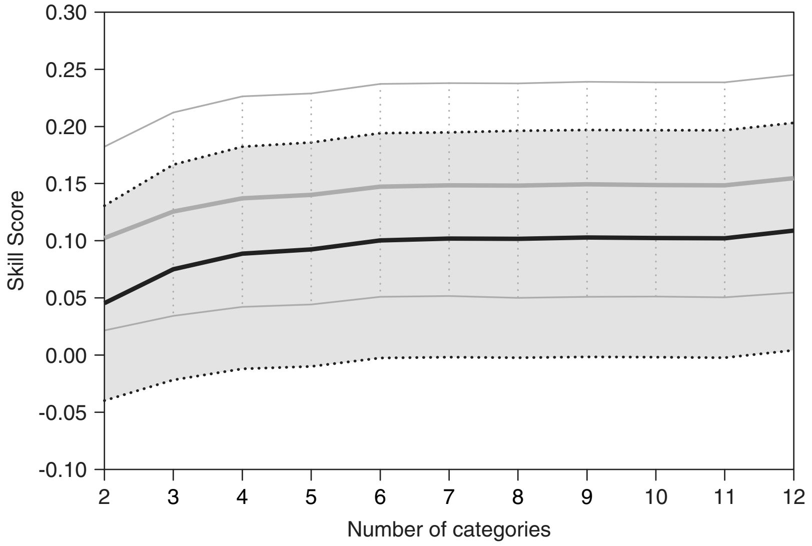

Figure 4. Brier skill score (BSS) with respect to climatology (black solid line) and expected for an infinite ensemble size (grey solid line) as a function of the number of probability categories for simplified forecasts of an eleven-member ensemble. The results are for one-month lead seasonal forecasts of summer temperature above the upper tercile over the tropics (20°N–20°S) started on the 1st of May during the period 1981–2005. 90% confidence intervals are displayed with black dotted lines and grey solid lines for the sample BSS and the expected BSS of an infinitely large ensemble, respectively. For clarity purposes, the former interval has been shaded.

without simplification. For instance, the BSS obtained without simplification is 0.057 and 0.091 for ensembles of five and seven members, respectively, which is much lower than the BSS of 0.109 obtained for 11 ensemble members. As for the BSS for simplified probabilities, it ranges from 0.009 to 0.057 for five ensemble members and from 0.038 to 0.091 for seven ensemble members. The large decrease in BSS with smaller ensembles is mainly due to the degradation of the reliability when decreasing the size of the ensemble, while the resolution is much less sensitive to the change in ensemble size (not shown). For an ensemble size of nine or larger, both the BSS and the confidence intervals are similar regardless of the number of members, suggesting that at least in the case of the ECMWF system, the hindcast ensemble size is large enough to provide robust estimates of the BS and its associated terms. Another advantage of the larger ensembles is that the confidence intervals are narrower. On the basis of these results, a user might decide to discount the forecast quality estimates obtained with a set of ensemble hindcasts of a much smaller ensemble size than the size used for the forecasts.

A simplification has also been used in the past to compare the forecast quality of systems having different ensemble sizes (Mullen and Buizza, 2002; Hagedorn *et al*., 2005). If the same set of probability categories is used for both systems to estimate the BSS and its reliability and resolution terms, the results in Figure 3 suggest that the system with the largest ensemble size will be favoured in the case of small ensembles, regardless of the number of categories. As the BSS depends both on the ensemble size and the number of categories in the simplification of the probabilities, especially for small ensembles and number of categories, a comparison of the forecast quality of two forecast systems should take into account both effects. Given a forecast system with a specific ensemble size, Ferro (2007) has developed an analytical expression, which depends on both the sample BSS and the sharpness, to estimate a value of the BSS for any ensemble size. Figure 4 compares the categorized BSS for the 11-member hindcasts and estimated for an infinite ensemble size. There is a clear gain in skill by increasing the ensemble size, with differences much larger than expected from the results in Figure 3, although confidence intervals are large. The BSS for infinite ensemble size is affected by the simplification in a similar way as the BSS estimated from the sample, a consequence of the resolution term being the most affected by the simplification. We suggest that the expected skill for an infinite ensemble size should also be conveyed to the user for the sake of completeness.

4. Summary

This study has investigated the effect on forecast quality of the simplification of probability forecasts issued with an ensemble of simulations in the context of operational seasonal ensemble forecasts. Operational probability forecasts issued from a dynamical ensemble are usually simplified by using a small number of probability categories in order to, among other reasons, smooth out the issued probabilities, avoid having empty probability categories, or issue warnings for specific events using probability thresholds. Examples of forecast probability simplification from the ECMWF operational seasonal forecast system have been used to illustrate the reduction in forecast quality due to a decrease in the number of forecast categories. The probability forecasts used in the example have significant resolution but are significantly unreliable. The simplification of the probabilities has the effect of reducing the BSS with respect to the BSS obtained, by using all the probability values that can be issued with a given ensemble. In the example, there is a reduction in forecast quality of 56 and 17% for two and five categories, respectively. This is a substantial reduction for forecast systems such as those typical in long-range forecasting, for which the skill is already moderately low. The reduction in BSS is mainly due to a decrease in resolution, the reliability term being only slightly affected.

The results have shown that BSS estimates depend on both the simplification of the forecast probabilities and the ensemble size used to estimate the probabilities, especially when both are small. A reduction in the number of categories decreases the forecast resolution and only slightly affects the forecast reliability. On the contrary, forecast reliability improves with the ensemble size, while forecast resolution shows only small changes. This implies that estimates of BS, reliability and resolution obtained with hindcast ensembles of smaller size (e.g. 11 members) than the one actually used to produce operational forecasts (e.g. 41 members) and with a different set of categories have the potential to differ substantially from the actual forecast quality of the real-time probability forecasts. This problem also applies to the comparison of forecast quality estimates of different forecast systems.

Categories are intervals of probability, so that a representative probability for the category is required to obtain forecast quality estimates. The examples used a weighted mean of all the probabilities included in the category as the representative probability. A comparison of these results with forecast quality estimates obtained with the representative probability taken as the centre of the probability interval show some relevant differences. We found a smaller reliability term and, as a consequence, an improved BSS with respect to the BSS obtained with the probabilities without simplification. The change in reliability is due to a shift of the representative probability of the category in the reliability diagram towards/away from the diagonal with respect to the probability calculated as the average of the original probabilities; in particular, in our example of overconfident sharp forecasts the shift is towards the diagonal, which explains the improvement in the reliability term, but at the same time reduces the sharpness and, as such, is equivalent to a calibration of the probabilities. The resolution term does not change because the shift of the representative forecast probability is in the horizontal direction in the attribute diagram, so that the vertical distance to the climatological frequency line that measures the contributions to the resolution term remains unaltered. However, the impact of the simplification over the BS is always towards a decrease of the forecast quality when the number of categories is reduced, regardless of the definition used for the representative probability of the category. In addition, the choice of the representative probability at the centre of the interval is arbitrary and introduces unexplained variations in the reliability term (and hence in the BSS) as a function of the number of categories, to the point that there is no monotonic relation between the BSS score and the number of categories (the monotonic relation can be seen in Figure 2). Hence, the use of the representative probability as a weighted average of the individual probabilities is a better option than the centre of the interval to illustrate the impact of the number of categories on forecast quality, in spite of most operational centres using the second option to define the representative probability.

Users of a set of forecasts should be provided with BS, resolution and reliability estimates for the specific set of probabilities that they receive. In particular, for a small number of probability categories, the reduction in forecast quality can be large enough for the forecast quality of the issued forecasts to be no longer significantly different from zero within sampling uncertainty. This happens in the example described here, if less than seven categories are used, and it implies that forecasters should consider how they should best simplify the probability forecasts when communicating them to the forecast user.

Acknowledgements

FJDR received financial support from the ENSEMBLES project (GOCE-CT-2003-505539). CASC was funded by Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP) under processes 2005/05210-7 and 2006/02497-6.

References

Anderson D, Stockdale T, Balmaseda M, Ferranti L, Vitart F, Molteni F, Doblas-Reyes FJ, Mogensen K, Vidard A. 2007. *Development of the ECMWF seasonal forecast System 3*. ECMWF Technical Memorandum 503 [Available from http://www.ecmwf.int/publications/library/do/references/show?id = 87744].

Atger F. 2004. Estimation of the reliability of ensemble-based probabilistic forecasts. *Quarterly Journal of the Royal Meteorological Society* **130**: 627–646.

Balmaseda M, Vidard A, Anderson D. 2007. *The ECMWF System 3 ocean analysis system*. ECMWF Technical Memorandum 508 [Available from http://www.ecmwf.int/publications/library/do/references/show?id=87667].

Barnston AJ, Mason SJ, Goddard L, Dewitt DG, Zebiak SE. 2003. Multimodel ensembling in seasonal climate forecasting at IRI. *Bulletin of the American Meteorological Society* **84**: 1783–1796.

Brier GW. 1950. Verification of forecasts expressed in terms of probabilities. *Monthly Weather Review* **78**: 1–3.

Ferro CAT. 2007. Comparing probabilistic forecasting systems with the Brier score. *Weather and Forecasting* **22**: 1076–1088.

Hagedorn R, Doblas-Reyes FJ, Palmer TN. 2005. The rationale behind the success of multi-model ensembles in seasonal forecasting. Part I: Basic concept. *Tellus Series A-Dynamic Meteorology and Oceanography* **57**: 219–233.

Hamill TM, Whitaker JS. 2006. Probabilistic quantitative precipitation forecasts based on re-forecast analogs: theory and application. *Monthly Weather Review* **134**: 3209–3229.

Hsu WR, Murphy AH. 1986. The attributes diagram: a geometrical framework for assessing the quality of probability forecasts. *International Journal of Forecasting* **2**: 285–293.

Jolliffe IT. 2007. Uncertainty and inference for verification measures. *Weather and Forecasting* **22**: 137–150.

Lanzante JR. 2005. A cautionary note on the use of error bars. *Journal of Climate* **18**: 3699–3703.

Mullen SL, Buizza R. 2002. The impact of horizontal resolution and ensemble size on probabilistic forecasts of precipitation by the ECMWF ensemble prediction system. *Weather and Forecasting* **17**: 173–191.

Murphy AH. 1986. A new decomposition of the Brier Score: formulation and interpretation. *Monthly Weather Review* **114**: 2671–2673.

Nicholls N. 2001. The insignificance of significance testing. *Bulletin of the American Meteorological Society* **82**: 981–986.

Richardson DS. 2001. Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quarterly Journal of the Royal Meteorological Society* **127**: 2473–2489.

Roulston MS, Smith LS. 2003. Combining dynamical and statistical ensembles. *Tellus Series A-Dynamic Meteorology and Oceanography* **55**: 16–30.

Stephenson DB, Coelho CAS, Jolliffe IT. 2008. Two extra components in the Brier score decomposition. *Weather and Forecasting* in press.

Stephenson DB, Coelho CAS, Balmaseda M, Doblas-Reyes FJ. 2005. Forecast Assimilation: a unified framework for the combination of multi-model weather and climate predictions. *Tellus Series A-Dynamic Meteorology and Oceanography* **57**: 253–264.

Uppala SM, Kallberg PW, Simmons AJ, Andrae U, da Costa Bechtold V, Fiorino M, Gibson JK, Haseler J, Hernandez A, Kelly GA, Li X, Onogi K, Saarinen S, Sokka N, Allan RP, Andersson E, Arpe K, Balmaseda MA, Beljaars ACM, van de Berg L, Bidlot J, Bormann N, Caires S, Chevallier F, Dethof A, Dragosavac M, Fisher M, Fuentes M, Hagemann S, Holm E, Hoskins BJ, Isaksen L, Janssen PAEM, Jenne R, McNally AP, Mahfouf JF, Morcrette JJ, Rayner NA, Saunders RW, Simon P, Sterl A, Trenberth KE, Untch A, Vasiljevic D, Viterbo P, Woollen J. 2005. The ERA-40 reanalysis. *Quarterly Journal of the Royal Meteorological Society* **131**: 2961–3012.

DOI: 10.1002/met