QUARTERLY JOURNAL OF THE ROYAL METEOROLOGICAL SOCIETY
Q. J. R. Meteorol. Soc. **136**: 1364–1370 (2010)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/qj.641

Decomposition of the Brier score for weighted forecast-verification pairs

R. M. B. Young*
Atmospheric, Oceanic and Planetary Physics, Department of Physics, University of Oxford, UK

Abstract: The Brier score is widely used in meteorology for quantifying probability forecast quality. The score can be decomposed into terms representing different aspects of forecast quality, but this implicitly requires each forecast-verification pair to be allocated equal weight. In this note an expression is derived for the decomposed Brier score which accounts for weighted forecast-verification pairs. A comparison of the unweighted and weighted cases using seasonal forecasts from the ENSEMBLES project shows that when weights are assigned proportional to the area represented by each grid point (weighting by cosine of latitude), the weighted forecasts give improved Brier and reliability scores compared with the unweighted case. This result is consistent with what is expected given that tropical predictability is generally better than extratropical predictability.

KEY WORDS forecast reliability; forecast resolution; numerical weather prediction; observational uncertainty; probability forecasting; seasonal forecasting

Received 19 December 2008; Revised 13 April 2010; Accepted 26 April 2010.

1 Introduction

The Brier (1950) score is widely used in meteorology for scoring probability forecasts with two mutually-exclusive outcomes (e.g. yes rain / no rain). For N forecast-verification pairs, the Brier score is†

$$\mathrm{BS} = \frac{1}{N}\sum_{n=1}^{N}(p_n - d_n)^2 \tag{1}$$

where p_n is the forecast probability for the first of the two outcomes to occur at verification point n, and

$$d_n = \begin{cases} 1 & \text{First outcome occurs} \\ 0 & \text{Second outcome occurs} \end{cases} \tag{2}$$

The expression in Eq. (1) implicitly assumes each of the forecast-verification pairs will be assigned equal weight. In some situations this is not appropriate, however, and each pair should be weighted accordingly. The weighted Brier score is

$$\mathrm{BS} = \frac{\sum_{n=1}^{N} w_n(p_n - d_n)^2}{\sum_{n=1}^{N} w_n} \tag{3}$$

where w_n is the weight assigned to forecast-verification pair n, and in this note general w_n is assumed. Murphy (1972, 1973) derived a decomposition of the non-weighted Brier score that splits it into terms representing observational uncertainty, forecast reliability, and forecast resolution. This decomposition is in common use, for example when using the attributes diagram (Hsu and Murphy, 1986). In this note an analogous decomposition is derived for weighted forecast-verification pairs. Hersbach (2000) has derived the equivalent weighted decomposition for the continuous ranked probability score.

2 When is weighting appropriate?

The decomposition derived in this note should only be applied in situations where weighting is suitable or necessary. While the decomposition applies for any well-defined weighting of forecast-verification pairs, it is useful to consider which situations are suitable for weighting and which are not.

A common situation which might require weighting is when the pairs are distributed non-uniformly in space, and a score is required which is representative of the whole domain. This could be over a regular grid such as the latitude-longitude grid (Fig. 1, top), where weighting each grid point by the cosine of latitude would approximate spatial integration (Jung and Leutbecher, 2008, for example). If there is some latitudinal variation in the Brier score or its components, then the effect of weighting can be substantial. Or it could be over an irregular set of

*Correspondence to: Atmospheric, Oceanic and Planetary Physics, Clarendon Laboratory, Parks Road, Oxford OX1 3PU, UK. Email: young@atm.ox.ac.uk. URL: http://www.atm.ox.ac.uk/user/young/

†Technically this is the half-Brier score as it only considers one of the two possible outcomes, but the factor of two can be omitted here without loss of generality.

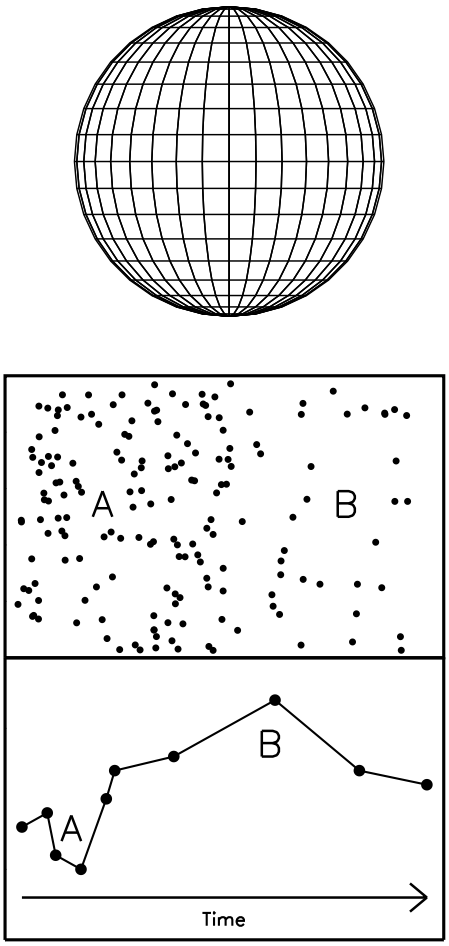

Figure 1. Three situations where weighting might be appropriate. Top: the latitude-longitude grid; for an average score it might be appropriate to weight the score for each grid element by the cosine of latitude. Middle: a set of irregularly-spaced points over a domain. If a Brier score is required that is representative of the whole domain then it might be appropriate to assign the points near B more weight than those near A, as the points near B are generally representative of larger areas. Bottom: a similar situation for weighting in time. If the score is representative of the whole time series then it might be appropriate to assign more weight to the points near B, as each one represents a longer period of time.

points such as a network of weather stations (Fig. 1, middle), where each station is representative of a different area. Alternatively, there might be a degree of redundancy between the observation stations in the more densely observed areas, so it might be appropriate to assign lower weight there.

Another situation that might require weighting is a series of forecast-verification pairs over time at a single point in space (Fig. 1, bottom). If the score represents the entire time series, then it might be appropriate to weight each pair by the length of the segment it represents. Alternatively, if the quality or reliability of observations changes over time (earlier verification samples may be less reliable or accurate because they used less advanced measurement techniques or a sparser observation network, for example), then it might be appropriate to assign less weight to the pairs with lower-quality observations. If specific regions of the domain are known to produce consistently unreliable observational data, then weighting might be appropriate even when the pairs are uniformly spaced.

There are also situations where the forecast-verification pairs are not distributed uniformly yet weighting is not appropriate. For example, a domain with very different climatological situations in two regions: a mountainous region with highly complex climate and many observation stations, and a plain region with few observation stations and a homogeneous climate. In this situation it would probably not be appropriate to assign higher weight to the observations on the plain just because each one is representative of a larger area, because the more complex climate in the mountains means the degree of redundancy between observations in the two regions may be comparable. Another example is when calculating the economic value of a forecast based on a cost/loss analysis (Wilks, 2001). Although the value score represents the whole domain it is based on total costs and total losses, which are just the sums of the costs and losses at individual points.

Overall, whether to weight or not depends on context. Whether the observations should be weighted is just as important in the interpretation of the results as the scores themselves. The derivation below assumes nothing about the weights themselves, however, so it is applicable in any situation where weighting is applied in practice.

3 Derivation of the weighted decomposition

Begin by defining

$$W = \sum_{n=1}^{N} w_n \tag{4}$$

as the total weight for the N pairs. The w_n could be normalized (so that $W = 1$ or $W = N$, for example), but W is retained here for generality. Following Murphy (1972), assume the forecast probability p_n can take any one of a fixed number of values; ordinarily these p_n are determined by the size of the forecast ensemble. For M ensemble members, p_n can take one of $T = M + 1$ values

$$p^t = \frac{t-1}{M} \qquad t \in \{1, 2, ..., M+1\} \tag{5}$$

where $t - 1$ is the number of ensemble members that predict the first of the two mutually exclusive outcomes will occur (Eq. 2). The Brier score can therefore be split into T categories BS^t, $t \in \{1, 2, ..., T\}$, each concerning the N^t cases whose forecast probability is p^t. Since by construction $p_n = p^t$ for all n in category t, then

$$\mathrm{BS}^t = \frac{\sum_{n=1}^{N^t} w_n^t (p^t - d_n^t)^2}{\sum_{n=1}^{N^t} w_n^t} \tag{6}$$

where the w_n^t are the weights assigned to the N^t forecast-verification pairs with $p_n = p^t$, and d_n^t is the outcome for the nth pair in category t.

Now define a second sum

$$w^t = \sum_{n=1}^{N^t} w_n^t \tag{7}$$

as the total weight in category t. Expand Eq. (6) and substitute in from Eq. (7) to get

$$\mathrm{BS}^t = (p^t)^2 - \frac{2p^t}{w^t} \sum_{n=1}^{N^t} w_n^t d_n^t + \frac{1}{w^t} \sum_{n=1}^{N^t} w_n^t (d_n^t)^2 \tag{8}$$

The final $(d_n^t)^2$ can be rewritten as d_n^t, because d_n^t can only take values of one and zero. Hence

$$\mathrm{BS}^t = (p^t)^2 - \frac{2p^t - 1}{w^t} \sum_{n=1}^{N^t} w_n^t d_n^t \tag{9}$$

Define a third quantity

$$\bar{d}^t = \frac{1}{w^t} \sum_{n=1}^{N^t} w_n^t d_n^t \tag{10}$$

which is the (weighted) relative frequency of the first outcome for forecasts in category t. Hence

$$\begin{aligned} \mathrm{BS}^t &= (p^t)^2 - (2p^t - 1)\bar{d}^t & (11)\\ &= (p^t - \bar{d}^t)^2 + \bar{d}^t(1 - \bar{d}^t) & (12) \end{aligned}$$

by completing the square. This form is analogous to Eq. (3) in Murphy (1972) but with $\bar{d}^t$ defined for weighted forecasts.

Now sum over the probability categories $1 \to T$ to recover the full Brier score. Each BS^t is weighted by the total weight in category t. Hence

$$\mathrm{BS} = \frac{1}{W} \sum_{t=1}^{T} w^t \mathrm{BS}^t \tag{13}$$

Substituting in from Eq. (12) gives

$$\mathrm{BS} = \frac{1}{W} \sum_{t=1}^{T} w^t \left(p^t - \bar{d}^t\right)^2 + \frac{1}{W} \sum_{t=1}^{T} w^t \bar{d}^t (1 - \bar{d}^t) \tag{14}$$

which is the 'original partition' of the Brier score defined by Murphy (1973, his Eq. 2) but for weighted forecasts. The first term on the r.h.s. represents the contribution to the Brier score due to the *forecast reliability*, denoted REL.

Now proceed similarly to Murphy (1973). Expanding the second term on the r.h.s. of Eq. (14):

$$\mathrm{BS} = \mathrm{REL} + \frac{1}{W} \sum_{t=1}^{T} w^t \bar{d}^t - \frac{1}{W} \sum_{t=1}^{T} w^t (\bar{d}^t)^2 \tag{15}$$

and define a fourth sum as the second term on the r.h.s. of the equation above:

$$\bar{d} = \frac{1}{W} \sum_{t=1}^{T} w^t \bar{d}^t \tag{16}$$

In Murphy (1973), $\bar{d}$ is the overall observed relative frequency of the first outcome. It can be shown that Eq. (16) is the equivalent for the weighted decomposition by substituting in for $\bar{d}^t$ from Eq. (10), giving:

$$\bar{d} = \frac{1}{W} \sum_{n=1}^{N} w_n d_n \tag{17}$$

which is just the weighted overall relative frequency of the first outcome. Hence

$$\begin{aligned} \mathrm{BS} &= \mathrm{REL} + \bar{d} - \frac{1}{W} \sum_{t=1}^{T} w^t (\bar{d}^t)^2 & (18)\\ &= \mathrm{REL} + \bar{d} - \bar{d}^2 \\ &\quad - \left\{ \frac{1}{W} \sum_{t=1}^{T} w^t (\bar{d}^t)^2 - 2\bar{d}^2 + \bar{d}^2 \right\} & (19)\\ &= \mathrm{REL} + \bar{d}(1 - \bar{d}) - \{\cdots\} & (20) \end{aligned}$$

The second term on the r.h.s. is the *observational uncertainty* for weighted forecasts, denoted UNC. Finally consider the term in braces; by expressing the second and third parts as sums and using Eqns (4) and (16), it is straightforward to show that this term is

$$\mathrm{RES} \equiv \frac{1}{W} \sum_{t=1}^{T} w^t \left(\bar{d}^t - \bar{d}\right)^2 \tag{21}$$

which is the *forecast resolution* term. By putting these together the decomposed Brier score for weighted forecast-verification pairs is obtained:

$$\begin{aligned} \mathrm{BS} &= \mathrm{UNC} + \mathrm{REL} - \mathrm{RES} & (22)\\ &= \bar{d}(1 - \bar{d}) \\ &\quad + \frac{1}{W} \sum_{t=1}^{T} w^t \left(p^t - \bar{d}^t\right)^2 \\ &\quad - \frac{1}{W} \sum_{t=1}^{T} w^t \left(\bar{d}^t - \bar{d}\right)^2 & (23) \end{aligned}$$

where W is given by Eq. (4), p^t by Eq. (5), w^t by Eq. (7), $\bar{d}^t$ by Eq. (10), and $\bar{d}$ by Eq. (17).

4 An illustration using seasonal forecasts

The effect of weighting on the Brier score and its decomposition is illustrated in this section with an analysis of seasonal forecasts from the EU ENSEMBLES project (Hewitt, 2005; Doblas-Reyes *et al.*, 2009). To illustrate the method, a weighting function is used that has predictable results: seasonal predictability is generally higher

at tropical latitudes compared with extratropical latitudes, so a weighting that favours the lower latitudes should produce better verification scores than the unweighted case.

The forecasts used are Stream 2 forecasts from five of the models used in the project: ECMWF IFS/HOPE, MeteoFrance ARPEGE4/OPA, Met Office HadGEM2, INGV ECHAM5/OPA8.2, and Kiel ECHAM5/OM1[‡]. Many datasets are available from these forecasts, so the analysis is restricted to the following: Each forecast consists of nine initial condition ensemble members started from 1 May for each of the years 1991–2001, and the quantity predicted is the monthly mean temperature 2 m above the surface for each of the subsequent seven months (i.e. a lead time of one month represents the mean from 1 May to 31 May). The forecasts were re-gridded from their original model grids[§] using the Climate Data Operators first order conservative remapping command *remapcon* onto a regular latitude-longitude grid of 2.5° spacing in both directions. They are verified grid point-wise against the ERA-40 reanalysis dataset (Uppala *et al.*, 2005)[¶] valid at the same locations, which gives $N = 10512$ for the total number of forecast-verification pairs at each lead time.

For the purpose of this example, a suitable event must be defined for the forecasts to predict. The event to predict is as follows: The monthly mean 2 m temperature will be above the climate mean 2 m temperature, where the climate mean 2 m temperature is defined using the ERA-40 dataset as the mean of this quantity over the period 1961–1990 for each grid point and month.

For each forecast (i.e. for each of the models, for each of the eleven years forecast), the nine ensemble members were used to compute a probability forecast p_n for this event to occur at each grid point n. The probability is given by the number of ensemble members predicting a higher 2 m temperature than the climate mean, divided by the number of ensemble members. For each ensemble forecast this was done for each lead time, up to seven months ahead. The verification value d_n was then found by comparing the ERA-40 reanalysis value at that time with the climate mean for the appropriate month.

An example of one such probability forecast, showing p_n at each point, is shown in the top panel of Fig. 2 along with the verification d_n in the middle panel.

The Brier score for the forecast was then calculated using Eq. (3) as a global statistic over all the grid points, and the decomposition was calculated using Eq. (23). This was done for an unweighted forecast, setting $w_n = 1$ for each grid point, and for a weighted forecast. An appropriate weighting function is to assign weight to each grid point proportional to the area of the globe that the point represents. For grid points at latitude λ_j separated in latitude by $\Delta\lambda$, this weighting function is given by[‖]

$$w_n = \begin{cases} \cos\left[\lambda_j\right] \sin\left[\frac{1}{2}\Delta\lambda\right] & -90^\circ < \lambda_j < 90^\circ \\ \sin^2\left[\frac{1}{4}\Delta\lambda\right] & \lambda_j = \pm 90^\circ \end{cases} \quad (24)$$

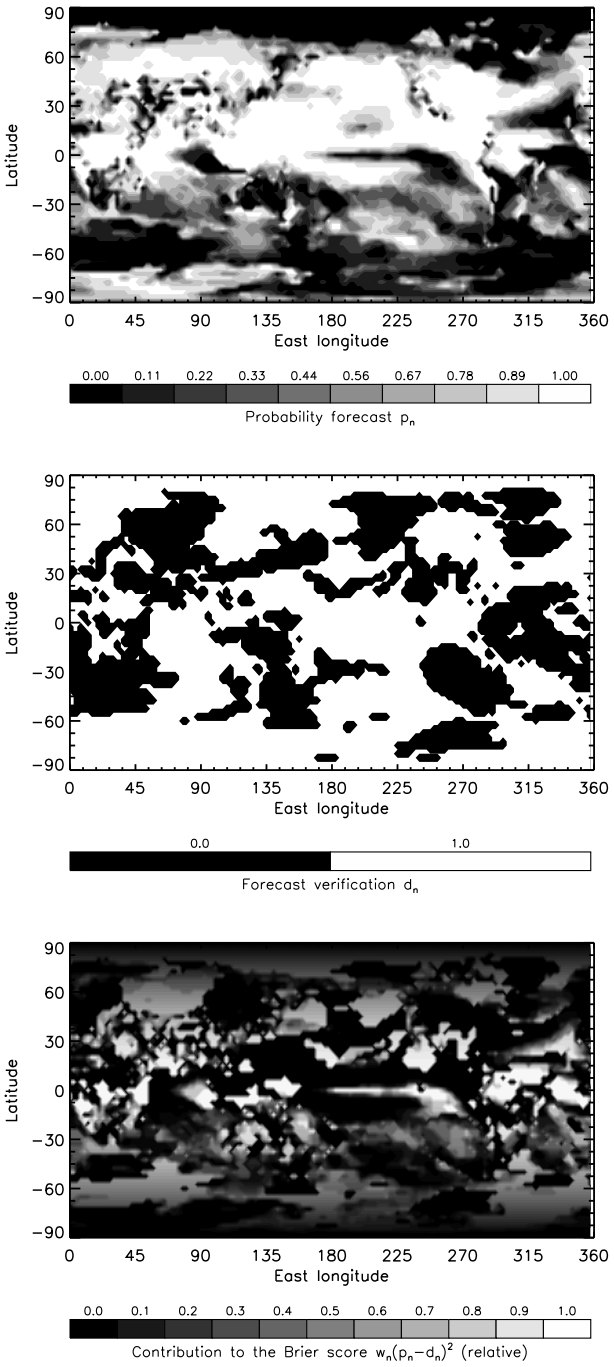

Figure 2. Example of a probability forecast of 2 m temperature used in the seasonal forecast analysis, showing results from the Met Office HadGEM2 forecast starting from 1 May 1991 and verifying against monthly means calculated between 1 August and 31 August (lead time four months). Top: Probability p_n assigned by the forecast at each grid point to the event occurring (monthly mean greater than climate mean for that month). Middle: Observed result d_n at each grid point. Bottom: Relative contributions $w_n(p_n - d_n)^2$ to the Brier score from each grid point for the weighted forecast, using the weights in Eq. (24).

[‡]Details of the experiments run as part of the ENSEMBLES project are listed at http://www.ecmwf.int/research/EU_projects/ENSEMBLES/table_experiments/

[§]N80 Gaussian reduced grid for ECMWF; regular long-lat grid of 192×145 points for the Met Office; N48 Gaussian grids for Kiel and INGV, and reduced 128x64 Gaussian grid for MeteoFrance.

[¶]http://www.ecmwf.int/products/data/archive/descriptions/e4/

where multiplicative factors constant at each grid point are omitted without loss of generality. This weighting clearly favours the low latitudes, and so better verification scores are expected than in the unweighted case, because seasonal predictability is generally higher at tropical latitudes compared with extratropical latitudes.

For the example forecast in Fig. 2, the contribution to the weighted Brier score from each point on the grid is shown in the bottom panel of that figure. For this particular forecast at lead time four months, the Brier score and its components are

	Unweighted	Weighted
Brier score	0.401	0.339
Uncertainty	0.210	0.227
Reliability	0.193	0.116
Resolution	0.00227	0.00392

The weighting function used in Eq. (24) assigns more weight at lower latitudes, so it is expected that the weighted forecasts will produce better verification scores than the unweighted case, because of the variation in seasonal predictability with latitude. The example shown in Fig. 2 confirms this prediction: in the table the Brier score and reliability component are lower for the weighted forecast, and the resolution component is higher.

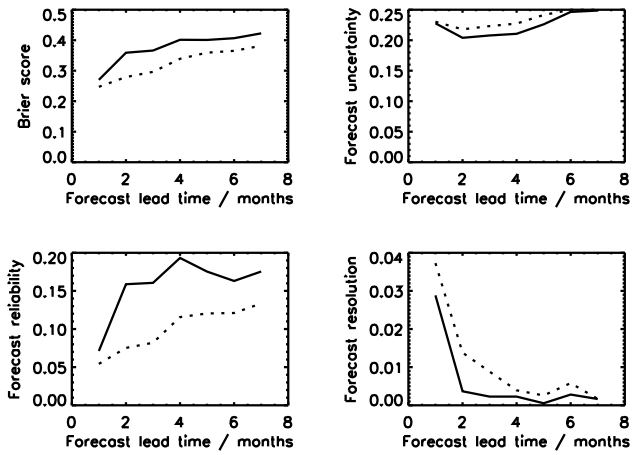

Figure 3. Brier score and its components for the Met Office HadGEM2 forecast as a function of lead time from 1 May 1991. From left to right, top to bottom: Brier score, uncertainty, reliability, and resolution components. In each plot the solid line is the unweighted case and the dotted line is the weighted case using the grid point weights in Eq. (24).

But is this a general result? Two more analyses are now presented: a single forecast verified over several months, and all the models and years available for analysis in combination. First, in Fig. 3, the scores for the weighted and unweighted cases are shown as a function of lead time for the Met Office HadGEM2 1991 forecast (the same forecast as Fig. 2). In this forecast the weighted forecast produces better scores than the unweighted forecast at all lead times, as predicted.

Second, the same analysis is extended to all the models and years for which results are available. The same calculations were done for all five models from 1991 to 2001. For ease of visualisation, the results were then combined for each model into a mean at each lead time

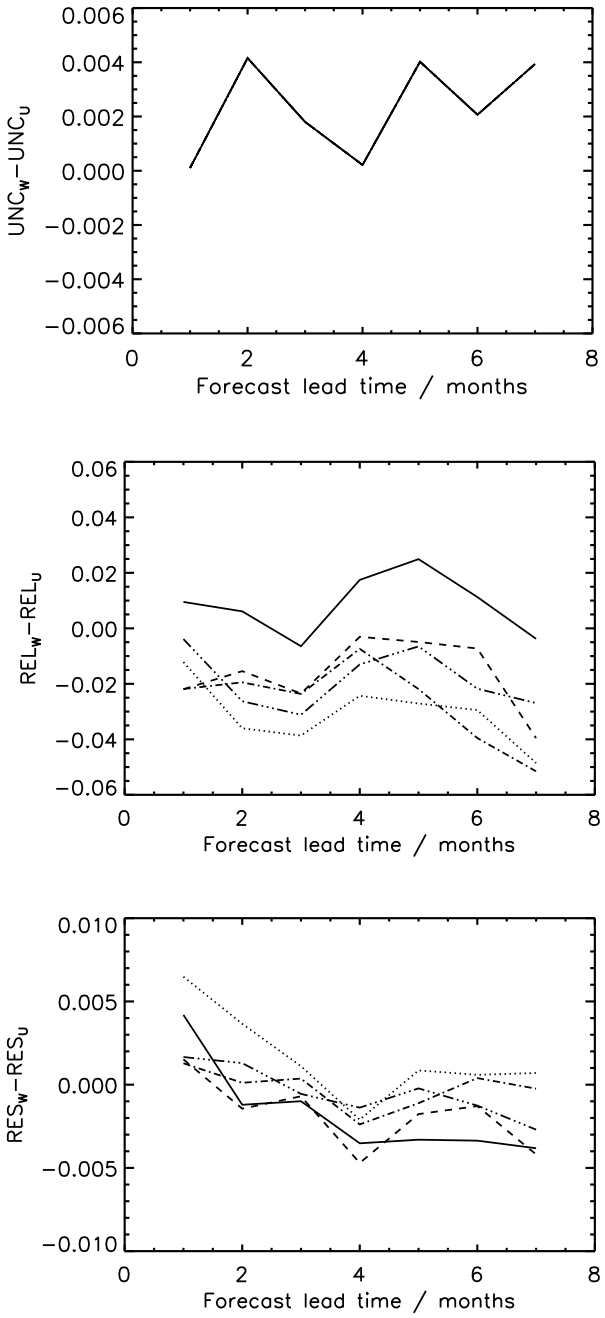

Figure 4. Difference in the Brier score components between the weighted and unweighted cases for all of the models and years considered. From top to bottom: uncertainty, reliability, and resolution components. Each line represents one model, and the value at each lead time is the difference between weighted and unweighted values of the mean score over all the years for which forecasts were analysed. The uncertainty appears as a single line as that quantity is a function of the observations only. The models are as follows: ECMWF IFS/HOPE (—), Met Office HadGEM2 (· · ·), Kiel ECHAM5/OM1 (– – –), INGV ECHAM5/OPA8 (– · –·), and MeteoFrance ARPEGE4/OPA (– · · ·).

over all the years analysed. These results are presented in Fig. 4. The differences between the weighted and unweighted cases for the Brier score (not shown in the figure) and the reliability component behave as predicted when latitudinal-based weighting is applied. The values in the weighted case are lower than in the unweighted case, as expected if the grid points where predictability is higher are favoured by the weighting scheme. The ECMWF model seems to behave in the opposite way, however, with the Brier score and reliability component becoming poorer when weighting is applied. Perhaps there is a bias in the ECMWF model that causes it to perform more favourably than the rest of the models at extratropical latitudes, or it might be because the ECMWF model is the same one used to create the ERA-40 dataset used for verification.

The differences in the resolution and uncertainty between unweighted and weighted cases have a smaller effect on the Brier score than the reliability. From Fig. 4, the relative contributions to the change in the Brier score when weighting is applied are approximately in the ratio 1:10:1 for uncertainty : reliability : resolution. Looking at the resolution scores, the predicted result is obtained for short lead times, as the resolution score increases with weighting. After three months lead time there is either no difference or a slight bias towards poorer scores when grid points are weighted. This may be partly because at lead times beyond two months resolution scores are close to zero anyway; see the bottom right hand panel of Fig. 3, for example.

The difference in the uncertainty scores is slightly anomalous, as the scores are marginally larger in the weighted case than in the unweighted case. As uncertainty can be interpreted as a measure of the intrinsic difficulty of the forecast, the uncertainty might be expected to decrease as more weight is assigned to regions where the behaviour is easier to predict. This result can be explained by examining the verification data, however. The definition of the event being forecast means that the expected value of $\bar{d}$ is 0.5 over the period covered by the climate mean. If the monthly mean temperature increases (decreases) after the end of the climate mean period, however, $\bar{d}$ will increase (decrease) because the probability of being greater than the climate mean rises above (falls below) 0.5. For both an increase and decrease the uncertainty will decrease, however, as it is equal to $\bar{d}(1-\bar{d})$. The amount $\bar{d}$ changes (and hence the amount the uncertainty falls) varies directly with the monthly mean temperature change. If the weighted uncertainty is larger than the unweighted value, therefore, the weighting must be assigning *more* weight to regions where temperature changes *less* between the climate mean period and the verification time. The result in Fig. 4 therefore predicts that the temperature over the 1991–2001 period has changed (with respect to the climate mean) more in the extratropics than in the tropics. In Fig. 5 this temperature anomaly is plotted for May, comparing the mean of the 1991–2001 monthly mean temperatures with the mean of the 1961–1990 period. The anomaly is greater in the extratropics than in the tropics, confirming the prediction and explaining the uncertainty results above. Equivalent plots for other months give the same result, and in some cases the difference between low and high latitudes is even greater than the example shown here.

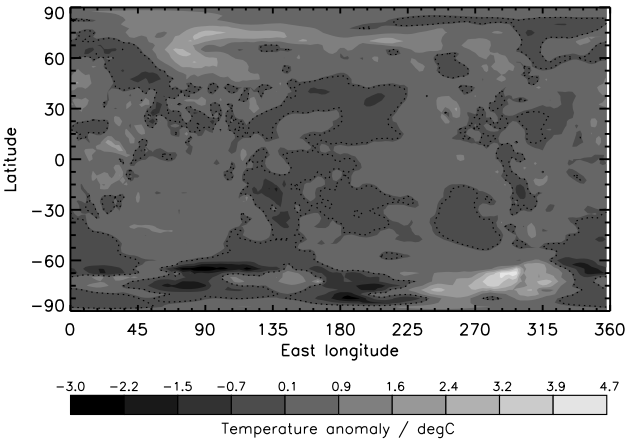

Figure 5. Temperature difference between the mean of monthly mean 2 m temperatures for May in the period 1991–2001 compared with the equivalent climate mean over 1961–1990. The dotted line is the zero contour.

5 Concluding remarks

In this note a decomposition of the Brier score has been derived for weighted forecast-verification pairs, and its use has been illustrated for seasonal forecasts weighted according to the area represented by each grid point, i.e. proportional to the cosine of latitude. The weighted forecasts in the example give improved Brier and reliability scores compared with the unweighted case, consistent with what is expected given that tropical predictability is generally better than extratropical predictability.

The new decomposition has a few consequences for other verification scores. The attributes diagram (Hsu and Murphy, 1986) plots forecast probability against observed relative frequency. For weighted pairs the ordinate on the attributes diagram should be changed to the weighted expression for $\bar{d}^t$ (Eq. 10), and the point size used to represent each forecast probability should be changed from the total number of observations in that category to the total weight in that category, w^t (Eq. 7).

The Brier score and its decomposition are often computed using a contingency table like Table 1 of Murphy and Winkler (1987). When weighted forecast-verification pairs are used, each element of the table is changed from the number of pairs with that combination of forecast and outcome to the total weight assigned to pairs with that combination. Equivalently, the contingency table for the whole forecast is a weighted sum of the contingency tables for each individual pair.

Weighting can also be applied to other scores. The ignorance score (Roulston and Smith, 2002) is defined by $\mathrm{IGN} = -\log_2 f_j$, where f_j is the forecast probability assigned to the observed outcome. Averaging over N

forecast-verification pairs with weights w_n gives

$$\langle \mathrm{IGN} \rangle = -\frac{1}{W}\sum_{n=1}^{N} w_n \log_2 f(n)_{j(n)} \quad (25)$$

$$= \sum_{n=1}^{N} \log_2 \left[f(n)_{j(n)}^{-w_n/W} \right] \quad (26)$$

where W is defined by Eq. (4). The quantities used to calculate points on the relative operating characteristic curve (Wilks, 2001, Fig. 4) are also affected by weighting: the hit rate and probability of false detection scores used to create the curve need to be calculated using the weight assigned to each element of the contingency table instead of the number of pairs. Finally, when constructing the rank histogram, using a set of weighted forecast-verification pairs means changing the value for each ensemble member bin from the number of pairs where the verification falls within that bin to the total weight of pairs falling within that bin.

Acknowledgements

This work was financially supported by NERC Studentship NER/S/A/2005/13667. The author thanks Peter Read, Martin Leutbecher, and two anonymous reviewers for their comments on the manuscript, and Falk Niehörster for comments on the manuscript and for access to the LSE Centre for the Analysis of Time Series seasonal forecasts from the ENSEMBLES project.

References

Brier GW. 1950. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **78**(1): 1–3, doi:10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2.

Doblas-Reyes FJ, Weisheimer A, Déqué M, Keenlyside N, McVean M, Murphy JM, Rogel P, Smith D, Palmer TN. 2009. Addressing model uncertainty in seasonal and annual dynamical ensemble forecasts. *Q. J. R. Meteorol. Soc.* **135**(643): 1538–1559, doi:10.1002/qj.464.

Hersbach H. 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting* **15**(5): 559–570, doi:10.1175/1520-0434(2000)015⟨0559:DOTCRP⟩2.0.CO;2.

Hewitt C. 2005. The ENSEMBLES Project. *EGU Newsletter* **13**: 22–25, URL http://www.the-eggs.org/data/eggs_13.pdf.

Hsu W, Murphy AH. 1986. The attributes diagram. A geometrical framework for assessing the quality of probability forecasts. *Int. J. Forecasting* **2**(3): 285–293, doi:10.1016/0169-2070(86)90048-8.

Jung T, Leutbecher M. 2008. Scale–dependent verification of ensemble forecasts. *Q. J. R. Meteorol. Soc.* **134**(633): 973–984, doi:10.1002/qj.255.

Murphy AH. 1972. Scalar and vector partitions of the probability score: Part I. Two-state situation. *J. Appl. Meteorol.* **11**(2): 273–282, doi:10.1175/1520-0450(1972)011⟨0273:SAVPOT⟩2.0.CO;2.

Murphy AH. 1973. A New Vector Partition of the Probability Score. *J. Appl. Meteorol.* **12**(4): 595–600, doi:10.1175/1520-0450(1973)012⟨0595:ANVPOT⟩2.0.CO;2.

Murphy AH, Winkler RL. 1987. A General Framework for Forecast Verification. *Mon. Weather Rev.* **115**: 1330–1338, doi:10.1175/1520-0493(1987)115$⟨$1330:AGFFFV$⟩$2.0.CO;2.

Roulston MS, Smith LA. 2002. Evaluating Probabilistic Forecasts Using Information Theory. *Mon. Weather Rev.* **130**(6): 1653–1660, doi:10.1175/1520-0493(2002)130⟨1653:EPFUIT⟩2.0.CO;2.

Uppala SM, KÅ llberg 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, Hólm 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 re-analysis. *Q. J. R. Meteorol. Soc.* **131**: 2961–3012, doi:10.1256/qj.04.176.

Wilks D. 2001. A skill score based on economic value for probability forecasts. *Meteorol. Appl.* **8**(2): 209–219, doi:10.1017/S1350482701002092.

