Evaluation of ensemble forecast uncertainty using a new proper score: Application to medium-range and seasonal forecasts

H. M. Christensen,a* I. M. Morozb and T. N. Palmera

aAtmospheric, Oceanic and Planetary Physics, University of Oxford, UK
bOxford Centre for Industrial and Applied Mathematics, University of Oxford, UK

*Correspondence to: H. M. Christensen (nee H. M. Arnold), Atmospheric, Oceanic and Planetary Physics, Clarendon Laboratory, Parks Road, Oxford, OX1 3PU, UK. E-mail: h.m.arnold@atm.ox.ac.uk

Forecast verification is important across scientific disciplines, as it provides a framework for evaluating the performance of a forecasting system. In the atmospheric sciences, probabilistic skill scores are often used for verification, as they provide a way of ranking the performance of different probabilistic forecasts unambiguously. In order to be useful, a skill score must be proper: it must encourage honesty in the forecaster and reward forecasts that are reliable and have good resolution. A new score, the error-spread score (ES), is proposed, which is particularly suitable for evaluation of ensemble forecasts. It is formulated with respect to the moments of the forecast. The ES is confirmed to be a proper score and is therefore sensitive to both resolution and reliability. The ES is tested on forecasts made using the Lorenz '96 system and found to be useful for summarizing the skill of the forecasts. The European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble prediction system (EPS) is evaluated using the ES. Its performance is compared with a perfect statistical probabilistic forecast: the ECMWF high-resolution deterministic forecast dressed with the observed error distribution. This generates a forecast that is perfectly reliable if considered over all time, but does not vary from day to day with the predictability of the atmospheric flow. The ES distinguishes between the dynamically reliable EPS forecasts and the statically reliable dressed deterministic forecasts. Other skill scores are tested and found to be comparatively insensitive to this desirable forecast quality. The ES is used to evaluate seasonal range ensemble forecasts made with the ECMWF System 4. The ensemble forecasts are found to be skilful when compared with climatological or persistence forecasts, though this skill is dependent on the region and time of year.

Key Words: error-spread score; forecast verification; reliability; uncertainty; proper scores; ensemble forecasting

Received 15 July 2013; Revised 17 March 2014; Accepted 25 March 2014; Published online in Wiley Online Library

1. Introduction: Evaluation of ensemble forecasts

It is well established that there are two major sources of uncertainty in weather forecasts: initial condition uncertainty and uncertainty due to the formulation of the weather prediction model (Ehrendorfer, 1997). Instead of producing a single deterministic forecast, an ensemble of forecasts should be generated in which the ensemble members explore these uncertainties (Palmer, 2001). The resultant probabilistic forecast is then issued to the user, who can make better informed decisions than if no uncertainty estimate were available. There are many different methods commonly used for verification of probabilistic forecasts. Graphical forecast verification techniques provide a comprehensive summary of the forecast, but it can be difficult to compare many forecast models using them. They can also be misleading, so must be interpreted carefully to draw the correct conclusions (Hamill, 2001). Instead, it is often necessary to choose a scalar summary of forecast performance which allows several forecasts to be ranked unambiguously. Scoring rules provide a framework for forecast verification. They summarize the accuracy of the forecast by giving a quantitative score based on the forecast probabilities and the actual outcome and can be considered as rewards that a forecaster wants to maximize.

Scoring rules must be carefully designed to encourage honesty from the forecaster: they must not contain features that promote exaggerated or understated probabilities. This constitutes a proper score, without which the forecaster may feel pressurized to present a forecast that is not their best guess (Brown, 1970). For example, a forecaster may want to be deliberately vague such that his or her prophecy will be proved correct, regardless of the outcome. Specifically, a proper scoring rule is one that is optimized if the true probability distribution is predicted, whereas a score is strictly proper if it is optimized if and only if the true distribution is predicted.

Ideally, the verification should behave like a random sample from the forecast probability distribution function (pdf)
events (e.g. rain or no rain). It is the mean square difference. The BS (Wilks, 2006) is used when considering dichotomous linear statistics (Goldstein and Wooff, 2007). Resolution is a property of the joint distribution of forecasts and observations: to have high resolution, a forecast must sort the observed states of the system into groups that are different from each other (Wilks, 2006; Leutbecher, 2010). A proper skill score must evaluate both the reliability and the resolution of a probabilistic forecast.

Currently, many scoring rules used for forecast verification, such as the continuous ranked probability score (CRPS: Wilks, 2006) and the ignorance score (IGN: Roulston and Smith, 2002), require an estimate of the full forecast pdf. This is usually achieved using kernel smoothing estimates or by fitting the parameters in some predetermined distribution, both of which require certain assumptions about the forecast pdf to be made. Alternatively, the pdf must be discretized in some way, such as for the Brier score (BS) and ranked probability score (RPS) (Wilks, 2006), which were both originally designed for multi-category forecasts. On the other hand, comparing the root-mean-square (RMS) error in the ensemble mean with the ensemble spread (Leutbecher and Palmer, 2008; Leutbecher, 2010) is an attractive verification tool, as it does not require an estimation of the full forecast pdf and instead is calculated using the raw ensemble forecast data. The RMS error and spread are displayed on scatter diagrams for subsamples of the forecast cases conditioned on the spread. This diagnoses the ability of the forecasting system to make flow-dependent uncertainty estimates. However, being a graphical diagnostic, it is very difficult to compare many forecast models using this tool.

We propose a new scoring rule designed for ensemble forecasts of continuous variables that is particularly sensitive to the reliability of a forecast and seeks to summarize the RMS error-spread scatter diagnostic. It will be formulated with respect to moments of the forecast distribution and not using the full distribution itself. These moments may be calculated directly from the ensemble forecast, provided it has sufficient members for an accurate estimate to be made. The new score we propose does not require the forecast to be discretized and acknowledges the inability of the forecaster to specify a probability distribution for a variable fully, due both to the amount of information needed to estimate the distribution and the number of bins needed to represent it on a computer. This limitation has been recognized by other authors and forms the basis for the development of Bayes linear statistics (Goldstein and Wooff, 2007).

The new score will be compared with a number of existing proper scores, detailed below.

### 1.1. The Brier score

The BS (Wilks, 2006) is used when considering dichotomous events (e.g. rain or no rain). It is the mean square difference between the forecast and observed probability of an event occurring:

\[
BS_i = (y_i - a_i)^2, \tag{1}
\]

where \(y_i\) is the \(i\)th predicted probability of the event occurring; \(a_i = 1\) if the event occurred and 0 otherwise. The BS must be evaluated for many test cases \(i\) and the arithmetic mean calculated.

The BS can be decomposed explicitly into its reliability and resolution components (Murphy, 1973). In particular, the reliability component of the BS is given by

\[
REL = \frac{1}{N} \sum_{p=1}^{P} n_p (\hat{f}_p - \bar{a}_p)^2, \tag{2}
\]

where \(N\) is the total number of forecast–observation pairs and \(P\) is the number of unique forecast probabilities issued; for example, \(P = 11\) if the allowed forecast probabilities are \([0.0, 0.1, 0.2, \ldots, 1.0]\). \(\hat{f}_p\) is the forecast probability, \(\bar{a}_p\) the observed frequency of the event given that \(\hat{f}_p\) was forecast and \(n_p\) is the number of times \(\hat{f}_p\) was forecast.

### 1.2. The ranked probability score

The RPS is used to evaluate a multi-category forecast. Such forecasts can take two forms: nominal, where there is no natural ordering of events, and ordinal, where the events are ordered numerically. For ordinal predictions, it is desirable that the score takes the ordering into account: a forecast should be rewarded for allocating high probabilities to events similar in magnitude to the observed event (a close distance in variable space). The RPS is defined as the squared sum of the difference between forecast and observed probabilities, so is closely related to the BS. However, in order to include the effects of distance discussed above, the difference is calculated between the cumulative forecast probabilities \(Y_{m,i}\) and the cumulative observations \(O_{m,i}\) (Wilks, 2006). Defining the number of event categories to be \(J\), the \(i\)th forecast probability for event category \(j\) to be \(y_{j,i}\) and the \(i\)th observed probability for event category \(i\) to be \(o_{j,i}\),

\[
Y_{m,j} = \sum_{j=1}^{m} y_{j,i}, \quad m = 1, 2, \ldots, J \tag{3}
\]

and

\[
O_{m,j} = \sum_{i=1}^{m} o_{j,i}, \quad m = 1, 2, \ldots, J, \tag{4}
\]

then

\[
RPS_i = \sum_{m=1}^{J} (Y_{m,i} - O_{m,i})^2 \tag{5}
\]

and the average RPS is calculated over many test cases \(i\). In this article, ten event categories are used, defined as the deciles of the climatological distribution.

### 1.3. Ignorance

IGN was proposed by Roulston and Smith (2002) as a way of evaluating a forecast based on the information it contains. As for the RPS, we define \(J\) event categories and consider \(N\) forecast–observation pairs. The forecast probability that the \(i\)th verification will be event \(j\) is defined to be \(f(i,j)\) (where \(j = 1, 2, \ldots, J\) and \(i = 1, 2, \ldots, N\)). If the corresponding outcome event was \(j(i)\), IGN is defined to be

\[
IGN_i = -\log_2 f(i,j(i)). \tag{6}
\]

IGN should be calculated by averaging over many forecast–verification pairs. As for the RPS, the deciles of the climatological distribution are used in this work to define ten categories for evaluating IGN. The expected value of IGN is closely related to the relative entropy of the forecast,

\[
R = \sum_i p_i \ln \left( \frac{p_i}{q_i} \right), \tag{7}
\]
where \( q_i \) is the climatological distribution and \( p_i \) is the forecast distribution (Roulston and Smith, 2002). Information theoretic measures, such as relative entropy and therefore IG, measure forecast utility and have been shown to be useful for quantifying predictability in chaotic systems (Kleeman, 2002; Majda et al., 2002).

1.4. **Skill scores**

Each of the above scores \( S \) may be converted into a skill score \( S_{ref} \) by comparison with the score evaluated for a reference forecast, \( S_{ref} \):

\[
SS = \frac{(S - S_{ref})}{(S_{ref} - S_{ref})},
\]

(8)

For all the scoring rules above the perfect score, \( S_{ref} \) is zero and the skill score can be expressed as

\[
SS = 1 - \frac{S}{S_{ref}}, \quad -\infty < SS \leq 1.
\]

(9)

2. **The error-spread score**

Consider two distributions, \( Q(X) \) is the truth probability distribution function, which has moments mean \( \mu \), variance \( \sigma^2 \), skewness \( \gamma \) and kurtosis \( \beta \), defined in the usual way:

\[
\mu = E[X],
\]

(10)

\[
\sigma^2 = E[(X - \mu)^2],
\]

(11)

\[
\gamma = E\left[\left(\frac{X - \mu}{\sigma}\right)^3\right],
\]

(12)

\[
\beta = E\left[\left(\frac{X - \mu}{\sigma}\right)^4\right],
\]

(13)

where \( E[\cdot] \) denotes the expectation of the variable. The probabilistic forecast issued is denoted \( P(X) \), with mean \( m \), variance \( s^2 \), skewness \( g \) and kurtosis \( b \), defined in the same way. The perfect probabilistic forecast will have moments equal to those of the truth distribution: \( m = \mu, s^2 = \sigma^2, g = \gamma \) and \( b = \beta, \) etc.

The error-spread score (ES) is written

\[
ES_i = (s_i^2 - e_i^2 - c_i s g_i)^2,
\]

(14)

where the difference between the verification \( z_i \) and the ensemble mean \( m_i \) is the error in the ensemble mean,

\[
e_i = (m_i - z_i),
\]

(15)

and the verification \( z_i \) follows the truth probability distribution \( Q \). The ES is calculated by averaging over many forecast-verification pairs \( i \), both from different grid-point locations and from different starting dates. A smaller average value of the score indicates a better set of forecasts.

The first two terms in the square of the right-hand side of Eq. (14) are motivated by the spread-error relationship; for a reliable ensemble, it is expected that the ensemble variance, \( s^2 \), will give an estimate of the expected squared error in the ensemble mean, \( e^2 \) (Leutbecher and Palmer, 2008). However, with these two terms alone, the score is not proper.\(^*\) Consider the trial score,

\[
ES_{trial} = (s^2 - e^2)^2.
\]

(16)

It can be shown that the expected value of this score is not minimized by predicting the true moments \( m = \mu, s = \sigma \). In fact it is minimized by forecasting \( m = \mu + \gamma \sigma / 2 \) and \( s^2 = \sigma^2(1 + \gamma^2/4) \) (see Appendix A). The substitutions \( m \rightarrow m + gs/2 \) and \( s^2 \rightarrow s^2(1 + g^2/4) \) transform the trial score into the ES (Eq. (14)), which can be shown to be a proper score (section 3).

The third term in the ES can be understood as acknowledging that the full forecast pdf contains more information than is in the first two moments alone. This term depends on the forecast skewness, \( g \). Consider the case when the forecast distribution is positively skewed (Figure 1). If the observed error is smaller than the predicted spread, \( e^2 < s^2 \), the verification must fall in either section B or C in Figure 1. The skewed forecast distribution predicts that the verification is more likely to fall in B, so this case is rewarded by the scoring rule. If the observed error is larger than the predicted spread, \( e^2 > s^2 \), the verification must fall in either section A or section D in Figure 1. Now, the forecast pdf indicates that section D is the more likely of the two, so the scoring rule rewards a negative error \( e \).

The ES is a function of the first three moments only. This can be understood by considering Eq. (16). When expanded, the resultant polynomial is fourth-order in the verification \( z \). The coefficient of the \( z^4 \) term is unity, i.e. it is dependent on the fourth power of the verification only, so the forecaster cannot hedge his or her bets by altering the kurtosis of the forecast distribution. The first term with a non-constant coefficient is the \( z^3 \) term, indicating that skewness is the first moment of the true distribution that interacts with the forecaster’s prediction. The forecast skewness is therefore important and should appear in the proper score. If the score were based on higher powers, for example motivated from \( (s^d - e^d)^2 \), the highest-order moment required would be the \( (2n - 1) \)th moment.\(^1\)

3. **Propriety of the error-spread score**

A scoring rule must be proper in order to be a useful measure of forecast skill. The ES cannot be strictly proper, as it is only a function of the moments of the forecast distribution; a pdf with the same moments as the true pdf will score equally well. However, it is important to confirm that the ES is a proper score.

To test for propriety, we calculate the expected value of the score, assuming the verification follows the truth distribution

\(^*\)Note that the relationship between the average spread and average square error is valid for any distribution, regardless of skewness (Leutbecher and Palmer, 2008). However, skewness becomes important when the instantaneous values of spread and error are compared, as in the ES.

\(^1\)Scores based on the magnitude of the error were also considered, for example \( |(z| - |e|)^2 \), but a proper score could not be found.
from a gamma distribution with moments (mean. However, it is comparatively insensitive to the spread of and penalizes biased forecasts which have a systematic error in the moments is the most skilful. indicates that forecasting a gamma distribution with the ‘true’ forecast (the true gamma distribution). As expected, each score converted to skill scores by comparison with the climatological and spread of the forecast gamma distribution. The scores are not necessarily equal to those for the ‘true’ verification gamma distribution. This is repeated for 10 000 test cases and the skill scores averaged over all cases.

4. Testing the error-spread score using the gamma distribution To analyze the sensitivity of the score to properties of the forecast and to demonstrate in what way it differs from existing scores, the ES score is first tested using an idealized statistical forecast and to demonstrate in what way it differs from existing rules must be minimized when the truth distribution is forecast. Appendix B confirms that the truth distribution falls at a stationary point of the score, and that this stationary point is a minimum. Therefore the scoring rule is proper, though not strictly proper, and is optimized by issuing the truth distribution. Appendix B also includes a second test of propriety, from Bröcker (2009).

5. Testing the error-spread score: Evaluation of forecasts in the Lorenz ‘96 system

Experiments are carried out in the Lorenz (1996, hereafter L96) simplified model of the atmosphere. The model consists of two scales of variables: a ring of large-scale, low-frequency X variables coupled to many smaller scale high-frequency Y variables. The interaction between these scales is described by the set of coupled equations

\[
\frac{dX_k}{dt} = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F \quad (19a)
\]

\[
\frac{dY_j}{dt} = -cY_j(Y_{j+2} - Y_{j-1}) - Y_j + \frac{hc}{b} X_{\text{int}}(j-1)/|j+1| + 1 \quad (19b)
\]

where the variables have cyclic boundary conditions: \(X_{K+K} = X_K\) and \(Y_{J+J} = Y_J\). The interpretation of the parameters in these equations and the values used in this study are shown in Table 1. The scaling of the variables is such that one model time unit is equal to five atmospheric days, deduced by comparing the error doubling time of the model with that observed in the atmosphere (Lorenz, 1996).

This model provides an excellent test bed for techniques in atmospheric modelling, as a robust ‘truth’ can be defined to use as a verification. The full two-scale model is run and the resultant X time series defined as truth. A simplified model in which the Y variables are assumed unresolved is also run to

\[
E[\text{ES}] = \left[ (\alpha^2 - \mu^2) + (\mu - m)^2 - \beta(\mu - m) \right]^2
\]

\[
+ \sigma^2 \left[ 2(\mu - m) + (\sigma - s) \right]^2
\]

\[
+ \sigma^4 (\beta - \gamma^2 - 1).
\]
Table 1. Parameter settings used for the L96 system.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symbol</th>
<th>Setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. X variables</td>
<td>K</td>
<td>8</td>
</tr>
<tr>
<td>No. Y variables per X variable</td>
<td>J</td>
<td>32</td>
</tr>
<tr>
<td>Coupling constant</td>
<td>h</td>
<td>1</td>
</tr>
<tr>
<td>Forcing term</td>
<td>F</td>
<td>20</td>
</tr>
<tr>
<td>Spatial scale ratio</td>
<td>b</td>
<td>10</td>
</tr>
<tr>
<td>Time-scale ratio</td>
<td>c</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 2. Tunable parameters in the forecast model for the L96 system and their values fitted from the truth time series.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symbol</th>
<th>Setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>a0</td>
<td>$a_0$</td>
<td>0.359</td>
</tr>
<tr>
<td>a1</td>
<td>$a_1$</td>
<td>1.31</td>
</tr>
<tr>
<td>a2</td>
<td>$a_2$</td>
<td>-0.0146</td>
</tr>
<tr>
<td>a3</td>
<td>$a_3$</td>
<td>-0.00230</td>
</tr>
<tr>
<td>$\sigma_n$</td>
<td>$\sigma_n$</td>
<td>1.99</td>
</tr>
<tr>
<td>$\phi$</td>
<td>$\phi$</td>
<td>0.985</td>
</tr>
</tbody>
</table>

produce forecasts of the state of the X variables. The forecast model is given by Eq. (20): the effect of the unresolved Y variables on the X variables is represented by a cubic polynomial in X and a first-order autoregressive (AR(1)) additive stochastic term is included to represent the variability of the unresolved Y variables.

$$\frac{dX_k}{dt} = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F - U(X_k), \quad (20a)$$

$$U(X_k) = a_0 + a_1X_k + a_2X_k^2 + a_3X_k^3 + \epsilon_k(t), \quad (20b)$$

$$\epsilon_k(t) = \phi \epsilon_k(t - \Delta t) + \sigma_n(1 - \phi^2)^{0.5}z(t). \quad (20c)$$

The cubic fit constants $a_i$, the standard deviation of the noise term $\sigma_n$ and the autocorrelation of the noise term $\phi$ are determined from the truth time series and their fitted values are given in Table 2. The random number $z(t) \sim N(0, 1)$. For more details on the methodology and for experiments using more complicated stochastic forecast models, refer to Arnold et al. (2013). 40 member ensemble forecasts were initialized from perfect initial conditions and their forecast skill evaluated at a lead time of 0.9 model time unit (~4.5 atmospheric days). The tunable parameters in the forecast model, $\sigma_n$ and $\phi$, were varied and the forecasts for each parameter pair evaluated using three techniques. Figure 3(a) shows the graphical error-spread diagnostic. The forecast-verification pairs are binned according to the variance of the forecast. The average variance in each bin is plotted against the MS spread.

$$\sigma / \sigma_{\text{meas}}$$

$0.00$ $0.08$ $0.16$ $0.24$ $0.32$ $0.40$ $0.48$ $0.56$ $0.64$ $0.72$ $0.80$ $0.88$ $0.96$ $1.04$ $1.12$ $1.20$ $1.28$ $1.36$ $1.44$ $1.52$ $1.60$ $1.68$ $1.76$ $1.84$ $1.92$ $2.00$

$\phi$

$0.00$ $0.001$ $0.002$ $0.003$ $0.004$ $0.005$ $0.006$ $0.007$ $0.008$ $0.009$ $0.010$ $0.011$ $0.012$ $0.013$ $0.014$ $0.015$ $0.016$ $0.017$ $0.018$ $0.019$ $0.020$ $0.021$ $0.022$ $0.023$ $0.024$ $0.025$

Figure 3. (a) The mean square (MS) error-spread diagnostic, (b) the reliability component of the Brier score and (c) the error-spread skill score, evaluated for forecasts of the L96 system using an additive stochastic parametrization scheme. In each figure, moving left–right, the autocorrelation of the noise in the forecast model, $\phi$, increases. Moving bottom–top, the standard deviation of the noise, $\sigma_n$, increases. The individual panels in (a) correspond to different values of ($\phi, \sigma_n$). The bottom row of panels in (a) are blank because deterministic forecasts cannot be analyzed using the MS error-spread diagnostic: there is no forecast spread to condition the binning on.
the mean square error in each bin. For a reliable forecast system, these points should lie on the diagonal (Leutbecher and Palmer, 2008). Figure 3(b) shows the reliability component of the Brier score (Brier, 1950; Murphy, 1973), REL (Eq. (2)), where the ‘event’ was defined as ‘the \( T_{850} \) variable is in the top third of its climatological distribution’. Figure 3(c) shows the new error-spread skill score (ESS), which is calculated with respect to the climatological forecast. The difficulty of analyzing many forecasts using a graphical method can now be appreciated. Trends can easily be identified in Figure 3(a), but the best set of parameter settings is hard to identify. The stochastic forecasts with small-magnitude noise (low \( \sigma_n \)) are underdispersive. The error in the ensemble mean is systematically larger than the spread of the ensemble, i.e. they are overconfident. However, the stochastic parametrizations with very persistent, large-magnitude noise (large \( \sigma_n \), large \( \phi \)) are overdispersive. Figure 3(b) shows REL evaluated for each parameter setting, which is small for a reliable forecast. It scores highly those forecasts where the variance matches the mean square error, such that the points in (a) lie on the diagonal. The ESS is a proper score and is also sensitive to the resolution of the forecast. It rewards well-calibrated forecasts, but also those that have a small error. The peak of the ESS in Figure 3(c) is shifted down compared to REL and it penalizes the large \( \sigma_n \), large \( \phi \) models for the increase in error in their forecasts. The ESS has summarized the results in Figure 3(a) and has shown a sensitivity to both reliability and resolution, as required of a proper score.

### 6. Testing the error-spread score: Evaluation of medium-range forecasts from the integrated forecasting system

The ESS was tested using 10 day operational forecasts made with the European Centre for Medium-Range Weather Forecasts (ECMWF) Ensemble Prediction System (EPS). The EPS uses a spectral atmosphere model, the Integrated Forecasting System (IFS). Out to day 10, the EPS is operationally run with a horizontal triangular truncation of T639, with 62 vertical levels, and uses persisted sea-surface temperature (SST) anomalies instead of a dynamical ocean model. The 50 member ensemble samples initial condition uncertainty using perturbations derived from an ensemble of data assimilations (EDA: Isaksen et al., 2010), which are combined with perturbations from the leading singular vectors (Buizza et al., 2008).

The EPS system uses stochastic parametrizations to represent uncertainty in the forecast due to model deficiencies; the 50 ensemble members differ, as each uses a different seed for the stochastic parametrization schemes. Two stochastic parametrization schemes are used. The stochastically perturbed parametrization tendencies (SPPT) scheme (Palmer et al., 2009) addresses model uncertainty due to physical parametrization schemes; it perturbs the parametrized tendencies about the mean of the scattered points lies on the diagonal for each case considered. However, the spread of the forecast does not indicate the error in the ensemble mean. In contrast, the EPS forecasts...

---

2 The IFS is a spectral model and resolution is indicated by the wave number at which the model is truncated. For comparison, a spectral resolution of T639 corresponds to a reduced Gaussian grid of N320, or 30 km resolution, or a 0.28° latitude/longitude grid.
contain information about the expected error in the ensemble mean. The spread of the EPS is well calibrated, though at latitudes close to the Equator it is slightly underdispersive (Figure 5(b)). At a lead time of ten days, the RMS error between the deterministic forecast and the verification is higher than the RMS error in the ensemble mean for the lower resolution EPS. This difference is greatest at midlatitudes and can be observed in Figure 5(c).

The skill of the EPS and DD forecasts will now be evaluated using the ES and the results compared with the RPS and IGN. However, since the ES uses the moments of the forecast distribution to evaluate the skill of the ensemble forecast, it is important to confirm that the number of ensemble members is sufficient to estimate the value of these moments reliably. To estimate the minimum ensemble size required to use the ES, the EPS ensemble forecasts are sampled to produce forecasts where the number of members, M = [5, 10, 15, ..., 50]. The RPS, IGN and ES are evaluated for each sub-ensemble and the globally averaged skill shown as a function of ensemble size in Figure 6. As expected, the ES converges more slowly than IGN or RPS, but by 40 members the ES estimate of skill appears to have converged. The ES is therefore unsuitable for ensemble forecasts with fewer than 40 members, but can be used to evaluate the 50 member EPS and DD forecasts described above.

Figure 7 shows the skill of the EPS forecasts calculated using these three proper scores: ES, RPS and IGN. For each, the smaller the score, the better the forecast. The BS was also calculated, but the results were very similar to the RPS and so are not shown here. All scores agree that the skill of the EPS forecast is lower in midlatitudes than at ±20°. However, they disagree as to the skill of the forecast near the Equator. The RPS and IGN indicate a reduced skill at the Equator, whereas ES indicates a higher skill there than at midlatitudes. The cause of this difference is that at the Equator the climatological variability is much smaller than at midlatitudes, so the climatological deciles are closer together. This affects the scores conditioned on the climatological percentiles (RPS, IGN), which do not account for the spacing of the bins. At the Equator, even if the forecast mean and verification are separated by several bins, the magnitude of the error is actually small. The ES score is not conditioned on the climatological deciles of the climatology and so is not susceptible to this. Forecasts made close to the Equator have a small RMS error. This is rewarded by the ES close to the Equator have a small RMS error. This is rewarded by the ES score, whereas the RPS and IGN are not sensitive to the magnitude of the error. It is therefore more informative and easier to compare the different scores if the skill of the forecast with respect to some reference forecast is considered for each scoring rule.

Figure 8(a) shows the ESS, RPSS and IGNSS evaluated with respect to the skill of a latitudinally and longitudinally dependent climatological forecast, (S_{\text{clim}}):

\[ SS = 1 - \frac{S_{\text{EPS}}}{S_{\text{clim}}}. \]  

Since the categories for both IGN and RPS are defined to be the deciles of the climatological distribution, the IGN and RPS for the climatological forecast (10% probability for each bin) are constant and do not vary as a function of position or test case. In constrast,
the ES for the climatological forecast is strongly latitudinally dependent, as the ES is sensitive to the absolute magnitudes of the forecast distribution: an ensemble forecast with small spread and small forecast error will score better than a forecast with large spread and large error. Despite these differences, when each skill score is calculated, all three proper skill scores rank different latitudes similarly. All three skill scores indicate little or no skill for forecasts made near the Equator: at this long lead time of 10 days, the climatological forecast is more skillful than the EPS in this region. The scores indicate higher forecast skill in the extratropics and all three indicate that forecasts in the Northern Hemisphere are more skillful than those in the Southern Hemisphere. The ESS accentuates the variation in skill as a function of latitude, but ranks different latitudes similarly to RPSS and IGNSS.

A reliable forecast will include a flow-dependent estimate of uncertainty in the forecast. In some cases, the ensemble will stay tight, indicating high predictability, whereas in other cases the spread of the ensemble will rapidly increase with time, indicating low predictability. It is desirable that a skill score be sensitive to how well a forecast captures flow-dependent uncertainty. The dressed deterministic forecast described above is perfectly reliable by construction if all verification cases are averaged over, but it does not include any information about flow-dependent uncertainty, as the spread is constant for a given lead time across all start dates. In contrast, the EPS forecasts include information about flow-dependent uncertainty and the spread of the ensemble varies from case to case depending on the predictability of the atmosphere. Figure 8(b) shows skill scores for the EPS forecast calculated with reference to the dressed deterministic (DD) forecast, using three different proper scores: ESS, RPSS and IGNSS. In each case, the skill score SS is related to the score for the EPS, $S_{\text{EPS}}$, and for the DD, $S_{\text{DD}}$, by

$$SS = 1 - \frac{S_{\text{EPS}}}{S_{\text{DD}}} \quad (22)$$

The higher the skill score, the better the scoring rule is able to distinguish between the dynamic probabilistic forecast made using the EPS and the static statistical forecast made using the DD ensemble. Figure 8(b) indicates that the ESS is considerably more sensitive to this property of an ensemble than the other scores, though it still ranks the skill of different latitudes comparatively. All scores indicate that forecasts of T850 at the Equator are less skillful than at other latitudes: the ESS indicates there is forecast skill at these latitudes, though the other scores suggest little improvement over the climatological forecast: the skill scores are close to zero.

It has already been observed that the deterministic forecast has a larger RMS error than the mean of the EPS forecast. This will contribute to the poorer scores for the DD forecast compared with the EPS forecast. A harsher test of the ability of the skill scores to detect flow-dependent forecast uncertainty is to compare the EPS forecast with a forecast that dresses the EPS ensemble mean with the correct distribution of errors. This dressed ensemble mean (DEM) forecast differs from the EPS forecast only in that it has a fixed ensemble spread (perfect on average), whereas the EPS produces a dynamic, flow-dependent indication of forecast uncertainty. Figure 8(c) shows the skill of the EPS forecast calculated with respect to the DEM forecast. The ESS is able to detect the skill in the EPS forecast from the dynamic reliability of the ensemble. Near the Equator, the EPS forecast is consistently underdispersive, so has negative skill compared with the DEM ensemble, which has the correct spread on average: the skill previously observed when comparing the EPS and DD forecasts is due to the lower RMS error for the EPS forecasts at equatorial latitudes. The other skill scores indicate only a slight improvement of the EPS over the DEM: compared with the ESS, they are insensitive to the dynamic flow-dependent reliability of a forecast.

7. Application to seasonal forecasts

Having confirmed that the error-spread score is a proper score, sensitive to both reliability and resolution but particularly sensitive to the reliability of a forecast, the score is used to evaluate forecasts made with the ECMWF seasonal forecasting system, System 4. In System 4, the IFS has a horizontal resolution of T255 (~80 km grid) with 91 levels in the vertical. The IFS is coupled to the ocean model, Nucleus for European Modelling of the Ocean (NEMO), and a 50 member ensemble forecast is produced out to a lead time of seven months. The forecasts are initialized from 1 May and 1 November for the period 1981–2010.

Three regions are selected for this case study: the Niño 3.4 (N3.4) region is defined as 5° S–5° N, 120–170° W, the equatorial Indian Ocean (EqIO) region is defined as 10° S–10° N, 50–70° E and the North Pacific (NPac) region is defined as 30–50° N, 130–180° W. The monthly and areally averaged SST anomaly forecasts are calculated for a given region and compared with the analysis averaged over that region. The forecasts made with System 4 are compared with two reference forecasts. The climatological forecast is generated by calculating the mean, standard deviation and skewness of the areally averaged reanalysis SST for each region over the 30 year time period considered. This forecast is therefore perfectly reliable, though it has no resolution. A persistence forecast is also generated. The mean of the persistence forecast is set to the average reanalysis SST for the month prior to the start of the forecast (e.g. April for the May initialized forecasts). The mean is calculated separately for each year and analysis increments are calculated as the difference between the SST reanalysis and the starting SST. The standard deviation and skewness of the analysis increments are calculated and used for the persistence forecast.
Figure 9. RMS error-spread diagnostic for System 4 seasonal forecasts of SST initialized in (a)–(c) May and (d)–(f) November. Forecasts of the average SST over each season were considered and compared with reanalysis. The upright dark grey triangles are for the Niño 3.4 region, the inverted mid-grey triangles are for the Equatorial Indian Ocean region and the light grey circles are for the North Pacific region, where the regions are defined in the text. To increase the sample size for this diagnostic, the unaveraged fields of SST in each region were used instead of their regionally averaged value.

Figure 9 shows the RMS error-spread diagnostic for each region calculated for each season. The spread of the forecasts for each region gives a good indication of the expected error in the forecast. However, it is difficult to identify which region has the most skillful forecasts: the EqIO has the smallest error on average, but the forecast spread does not vary greatly from the climatological spread. In contrast, the errors in the forecasts for the N3.4 region and the NPac region are much larger, but the spread of the ensemble also has a greater degree of flow dependence.

Figure 10 shows the average ES score calculated for each region for the System 4 ensemble forecasts, the climatological forecast and the persistence forecast for the May and November start dates, respectively. Figure 10(a) and (b) shows that System 4 forecasts for the N3.4 region have a high ES. However, the climatological and persistence forecasts score very poorly and have considerably higher ES than System 4 at all lead times for the May initialized forecasts and at some lead times for the November initialized cases. This indicates that there is considerable skill in the System 4 forecasts in this region: they contain significant information about flow-dependent uncertainty, which is not contained in the climatological or persistence forecast. In the N3.4 region, SST is dominated by the ocean component of the El Niño Southern Oscillation (ENSO). ENSO has a large degree of interannual variability, so the climatological and persistence forecasts score comparatively poorly. On the other hand, the oscillation has a high degree of predictability, which the ES indicates is reliably captured by the seasonal forecasts. These results suggest that ENSO is represented well in the IFS.

The System 4 forecasts for the EqIO (panels (c) and (d)) have the lowest (best) ES out of all the regions considered for both start dates. However, this region shows little variability, so the climatological and persistence forecasts also score well. The ES indicates that the probabilistic forecasts of SST in the EqIO region for November–March are skilful, containing information about flow-dependent uncertainty. The forecast skill then decreases slightly in April and May. Forecasts initialized in May show decreasing skill until August, but high skill for September–November, scoring better than climatology for these months despite the long lead time. In contrast, the climatological and persistence forecasts score similarly across the year. An atmospheric teleconnection exists between the ENSO region and the Indian Ocean (Klein et al., 1999; Alexander et al., 2002; Ding and Li, 2012). The teleconnection is most active in September and October and observations suggest that the phase of ENSO results in predictable changes in EqIO SST, which persist from October through to March/April (Ding and Li, 2012). In the spring, a slight ‘persistence barrier’ is observed in the EqIO region, after which
directly linked to their spread: the ES scores a reliable forecast reference forecasts as indicated by Figure 10 can be seen to be of lead time for all regions and both start dates. The skill of the construction, the difference in their scores is due to resolution.

Figure 10 also indicates how the ES balances consideration of Figure 10 also indicates how the ES balances consideration of Figure 10 also indicates how the ES balances consideration of Figure 10 also indicates how the ES balances consideration of Figure 10 also indicates how the ES balances consistency with knowledge of this teleconnection. As well as the ability to capture the behaviour of ENSO (demonstrated in panels (a) and (b)), panels (c) and (d) indicate that the IFS is able to capture the ENSO–Indian Ocean teleconnection successfully, resulting in predictability in the EqIO region.

In the NPac region (panels (e) and (f)), variations in SST are much greater, but the System 4 ensemble is unable to forecast the observed variations. This results in a higher ES in this region. The climatological and persistence forecasts are also poorer than in the EqIO, due to the high variability. There is observational evidence for an ENSO–North Pacific atmospheric teleconnection (Alexander et al., 2002). This results in NPac SST anticorrelated to those in the N3.4 region. The NPac SST anomalies develop during the boreal summer, after which they remain fairly constant in magnitude from September through to May. The increased predictability in NPac SST due to this teleconnection is not reflected in the skill of the System 4 forecasts, according to the ES. This indicates that the System 4 model is not able to capture this mechanism skilfully. Improving the representation of the ENSO–North Pacific teleconnection could be an interesting area for future research in System 4.

Consideration of Figure 10 also indicates how the ES balances scoring reliability and resolution in a forecast. Since the climatological and persistence forecasts are perfectly reliable by construction, the difference in their scores is due to resolution. Figure 11 shows the spread of each reference forecast as a function of lead time for all regions and both start dates. The skill of the reference forecasts as indicated by Figure 10 can be seen to be directly linked to their spread: the ES scores a reliable forecast with narrow spread as better than a reliable forecast with large spread. The ES results shown in Figure 10(a) and (b) indicate that, in absolute terms, System 4 performs equally well in boreal summer and winter in the N3.4 region, so any difference in skill scores between the summer and winter seasons can be attributed to differences in skill of the climatological forecast. In particular, the reduced climatological ENSO variability in boreal spring, shown in Figure 11(b), results in a skilful climatological forecast for these months, with both small spread and error resulting in small values of the ES.

In summary, Figure 10 shows that the ES detects considerable skill in System 4 forecasts when compared with climatological or persistence forecasts, but that this skill is dependent on the region under consideration and the time of year. From knowledge of the behaviour of the ES, the observed skill in the forecasts can be attributed to skill in the spread of the ensemble forecast, which gives a reliable estimate of the uncertainty in the ensemble mean and varies according to the predictability of the atmospheric flow.

8. Conclusion

A new scoring rule, the error-spread score (ES), has been proposed, which is particularly suitable for verification of ensemble forecasts and is particularly sensitive to the reliability of the forecast. It depends only on the first three moments of the forecast distribution, so it is not necessary to estimate the full forecast pdf to evaluate the score. It is also not necessary to use event categories to discretize the forecast pdf, as is the case for the RPS. The score is shown to be a proper score.

The behaviour of the ES was tested using ensemble forecasts drawn from the gamma distribution and compared with the behaviour of the RPS and IGN. The ES was found to be more sensitive to ensemble spread than either RPS or IGN: the ES penalized forecasts with incorrect mean or standard deviation, whereas RPS and IGN were considerably more sensitive to the mean of the forecast pdf than to the spread. This indicates that the ES is more sensitive to forecast reliability than the other scores tested.

The ES was tested using forecasts made in the Lorenz ’96 system and was found to be sensitive to both reliability and resolution, as expected. The score was also tested using forecasts made with the ECMWF IFS. The score was used to test both EPS forecasts, which have a dynamic representation of model uncertainty, and a ‘dressed deterministic’ ensemble forecast, which does not have a flow-dependent probability distribution. The ES was able to detect significant skill in the EPS forecasts, due to their ability to predict flow-dependent uncertainty. Existing scores (RPS and IGN) were also used to evaluate the skill of these test cases, but were found to be comparatively insensitive to this desirable property of probabilistic forecasts.

The ES was used to evaluate the skill of seasonal forecasts made using the ECMWF System 4 model. The score indicates significant skill in the System 4 forecasts, as the ensemble is able to capture the flow-dependent uncertainty in the ensemble mean. The annual variation in skill indicates that the IFS is successfully capturing the behaviour of ENSO, as well as atmospheric teleconnection driven by ENSO, which results in predictability in Indian Ocean SST.

The results indicate that the ES is a useful forecast verification tool due to its ease of use, computational cheapness and sensitivity to desirable properties of ensemble forecasts.

Acknowledgements

The authors thank J. Bröcker, C. Ferro and M. Leutbecher for helpful discussion regarding scoring rules and A. Weisheimer for providing and advising on the seasonal forecasts. Thanks also go to the editors and two anonymous reviewers for their constructive comments and suggestions, which helped to improve this manuscript significantly. HMC is thankful to the Aspen Center for Physics and NSF Grant number 1066293 for hospitality.
during the writing of this article. The research of HMC was supported by a NERC studentship and the research of TNP was supported by ERC grant number 291406.

Appendix A: Derivation of the form of the error-spread score

The starting point when deriving the ES is the spread–error relationship; the expected squared error of the ensemble mean can be related to the expected ensemble variance by assuming that the ensemble members and the truth are independently identically distributed random variables with variance $\sigma^2$ (Leutbecher, 2010):

$$\frac{M}{M - 1} \text{ estimate ensemble variance} = \frac{M}{M + 1} \text{ squared ensemble mean error},$$  

(A1)

where the variance and mean error have been estimated by averaging over many forecast-verification pairs and $M$ is the size of the forecast ensemble. For large ensemble size, the correction factor is close to 1.

Consider the trial ES:

$$E[S_{\text{trial}}] = (s^2 - z^2)^2. \quad \text{(A2)}$$

Expanding out the brackets and expressing the error in terms of the forecast ensemble mean $m$ and the verification $z$, we have

$$E[S_{\text{trial}}] = (s - 2sz^2 + (m - z)^2 + (m - z)^4 \quad \text{(A3)}$$

The expected value of the score can be calculated by assuming the verification follows the truth distribution:

$$E[E[S_{\text{trial}}]] = (s^2 - 2sz^2 + m^2 + m^4) + E[z^2] (4m^2 - 2s^2 + 2m^2) - 4m E[z^2] + E[z^4]. \quad \text{(A4)}$$

The stationary points of the score can be calculated by differentiating with respect to the forecast moments:

$$F := \frac{d E[S_{\text{trial}}]}{ds} = 4s(s^2 - m^2) + 8m E[z] - 4s E[z^2], \quad \text{(A5)}$$

$$G := \frac{d E[S_{\text{trial}}]}{dm} = -4m(s^2 - m^2) + 4(s^2 - 12m^2) E[z] + 12m E[z^2] - 4E[z^4]. \quad \text{(A6)}$$

Substituting the true moments $E[z] = \mu$, $E[z^2] = \sigma^2 + \mu^2$ and $E[z^3] = \gamma \sigma^3 + 3\mu \sigma^2 + \mu^3$, we have

$$F = 4s(s^2 - \sigma^2 - (m - \mu)^2), \quad \text{(A7)}$$

$$G = 4(\sigma^2 - \mu^2) + 4(\sigma^2 - \mu^2) (3\sigma^2 - s^2) - 4\gamma \sigma^3. \quad \text{(A8)}$$

Setting $F = 0$ gives

$$s^2 = \sigma^2 + (m - \mu)^2. \quad \text{(A9)}$$

Setting $G = 0$ and substituting Eq. (A9) gives

$$4\gamma \sigma^3 = 4(\mu - \mu)^3 + 4(\sigma^2 - \mu^2) (3\sigma^2 - s^2) \quad \text{(A10)}$$

$$= 4(\mu - \mu)^3 (3\sigma^2 - \mu^2) + 2(\mu - \mu)^2 (3\sigma^2 - \sigma^2 - (m - \mu)^2) \quad \text{(B3)}$$

$$= 8\sigma^2 (m - \mu), \quad \text{(B4)}$$

$$\therefore \quad m = \mu + \frac{\gamma \sigma}{2}. \quad \text{(B5)}$$

Substituting Eq. (A10) into Eq. (A9) gives

$$s^2 = \sigma^2 + \left(\mu + \frac{\gamma \sigma}{2} - \mu\right)^2 \quad \text{(B6)}$$

$$= \sigma^2 \left(1 + \frac{\gamma^2}{4}\right). \quad \text{(B7)}$$

Therefore, the trial ES is not optimized if the mean and standard deviation of the true distribution are forecast. Instead of issuing his or her true belief, $(m, \sigma)$, the forecaster should predict a distribution with mean $m_{\text{hedged}} = m + gs/2$ and inflated standard deviation $s_{\text{hedged}} = s(1 + g^2/4)$ in order to maximize the expected score.

To prevent a forecaster from hedging the forecast in this way, the substitution $m \rightarrow m + gs/2$ and $s \rightarrow s(1 + g^2/4)$ can be made in the trial ES:

$$E[S_{\text{trial}}] := (s^2 - (m - z)^2)^2 \quad \text{(B8)}$$

$$\rightarrow \quad \text{ES} := (s^2 - (1 + \frac{g^2}{4} - (m + \frac{gs}{2} - z)^2)^2, \quad \text{(B9)}$$

$$\text{ES} = (s^2 - c^2 - cg)^2. \quad \text{(B10)}$$

Appendix B: Confirmation of propriety of the error-spread score

It is important to confirm that the ES is proper. Firstly, expand out the brackets:

$$E[S] = (s^2 - (m - z)^2 - (m - z)g)^2 \quad \text{(B11)}$$

$$= (s^2 - 2m^2 + 2ms^2 g + m^2 s^2 g^2 + 2m^3 s^2 g + m^4) + z(4ms^2 + 2s^3 g - 4m^2 s^2 g - 4m^3 - 2m^2 s^2 g - 2ms^3 g) + z^2(-2s^2 + 2m^2 + 2ms^2 + 4m^3 + 4ms^2 + s^2 g^2) + z^4(-4m - 2sg) + z^6. \quad \text{(B12)}$$

Calculate the expectation of the score assuming the verification, $z$, follows the truth distribution given by Eqs. (11)–(13):

$$E[E[z]] = (s^2 - 2m^2 s^2 + 2ms^2 g + m^2 s^2 g^2 + 2m^3 s^2 g + m^4) + E[z](4ms^2 + 2s^3 g - 4m^2 s^2 g - 4m^3 - 2m^2 s^2 g - 2ms^3 g) + E[z]^2(-2s^2 + 2m^2 + 2ms^2 + 4m^3 + 4ms^2 + s^2 g^2) + E[z]^4(-4m - 2sg) + E[z]^6. \quad \text{(B13)}$$

However,

$$E[z] = \mu, \quad \text{(B4)}$$

$$E[z^2] = \sigma^2 + \mu^2, \quad \text{(B5)}$$

$$E[z^3] = \gamma \sigma^3 + 3\mu \sigma^2 + \mu^3, \quad \text{(B6)}$$

$$E[z^4] = \sigma^4 \beta + 4\mu \sigma^2 \gamma + 6 \mu \sigma^2 \mu^4. \quad \text{(B7)}$$

Substituting these definitions, it can be shown that

$$E[E[z]] = [(\sigma^2 - s^2) + (\mu - m)^2 - \sigma g(\mu - m)^2] + \sigma^2 [2(\mu - m) + (\sigma \gamma - sg)]^2 + \sigma^4 (\beta - \gamma^2 - 1). \quad \text{(B14)}$$

In order to be proper, the expected value of the scoring rule must be minimized when the ‘truth’ distribution is forecast. Let us test this here.
Differentiating with respect to $m$:

$$\frac{dE[ES]}{dm} = 2\left[ (\sigma^2 - s^2) + (\mu - m)^2 - sg(\mu - m) \right] \times \left[ sg - 2(\mu - m) \right] - 4\sigma^2 \left[ 2(\mu - m) + (\sigma \gamma - sg) \right] \tag{B9}$$

$$= 0 \quad \text{at optimum.}$$

Differentiating with respect to $s$:

$$\frac{dE[ES]}{ds} = 2\left[ (\sigma^2 - s^2) + (\mu - m)^2 - sg(\mu - m) \right] \times \left[ -2s - g(\mu - m) \right] - 2\sigma^2 g \left[ 2(\mu - m) + (\sigma \gamma - sg) \right] \tag{B10}$$

$$= 0 \quad \text{at optimum.}$$

Differentiating with respect to $g$:

$$\frac{dE[ES]}{dg} = -2s(\mu - m) \times \left[ (\sigma^2 - s^2) + (\mu - m)^2 - sg(\mu - m) \right] - 2\sigma^2 g \left[ 2(\mu - m) + (\sigma \gamma - sg) \right] \tag{B11}$$

$$= 0 \quad \text{at optimum.}$$

Since

$$\frac{dE[ES]}{dv} = 0 \quad \text{for } v = m, s, g,$$

the 'truth' distribution corresponds to a stationary point of the score. The Hessian of the score is given by

$$H = 2\sigma^2 \begin{pmatrix} \gamma^2 + 4 & 0 & 2\sigma \\ 0 & \gamma^2 + 4 & 2\sigma \gamma \\ 2\sigma & 2\sigma \gamma & \sigma^2 \end{pmatrix},$$

which has three eigenvalues $\geq 0$. This stationary point is a minimum as required.

Additionally, a score, $S$ is proper if, for any two probability densities $P(x)$ and $Q(x)$ (Bröcker, 2009),

$$\int S[P(x), z] Q(z) dz \geq \int S[Q(x), z] Q(z) dz,$$  \tag{B12}

where the integral is over the possible verifications $z$. This criterion can be tested for the ES. The term on the left of Eq. (B12) is just the expectation of ES calculated earlier, if we identify $P(x)$ with the issued forecast and $Q(x)$ with the 'truth' distribution:

$$\int S[P(x), z] Q(z) = \left[ (\sigma^2 - s^2) + (\mu - m)^2 - sg(\mu - m) \right] + \sigma^2 \left[ 2(\mu - m) + (\sigma \gamma - sg) \right] + \sigma^4 (\beta - \gamma^2 - 1).$$ \tag{B13}

Similarly,

$$\int S[Q(x), z] Q(z) = \sigma^4 (\beta - \gamma^2 - 1).$$ \tag{B14}

Therefore,

$$\int S[P(x), z] Q(z) dz \geq \int S[Q(x), z] Q(z) = \left[ (\sigma^2 - s^2) + (\mu - m)^2 - sg(\mu - m) \right] + \sigma^2 \left[ 2(\mu - m) + (\sigma \gamma - sg) \right] + \sigma^4 (\beta - \gamma^2 - 1).$$ \tag{B15}

The error-spread score is a proper score.

References


