Received: 11 April 2018 | Revised: 20 July 2018 | Accepted: 20 November 2018 | Published on: 20 February 2019

DOI: 10.1002/qj.3447

SUPPLEMENT ARTICLE

Quarterly Journal of the Royal Meteorological Society RMetS

The ensemble-adjusted Ignorance Score for forecasts issued as normal distributions

Stefan Siegert[1] | Christopher A. T. Ferro[1] | David B. Stephenson[1] | Martin Leutbecher[2]

[1]Department of Mathematics, University of Exeter, UK
[2]ECMWF, Reading, UK

Correspondence
Stefan Siegert, Department of Mathematics, Harrison Building, Streatham Campus, University of Exeter, North Park Road, Exeter EX4 4QE, UK.
Email: s.siegert@exeter.ac.uk

Funding information
European Union Programme FP7/2007-13 under grant agreement 3038378 (SPECS),

This study considers the application of the Ignorance Score (IS, also known as the Logarithmic Score) for ensemble verification. In particular, we consider the case where an ensemble forecast is transformed to a normal forecast distribution, and this distribution is evaluated by the IS. It is shown that the IS systematically depends on the ensemble size, such that larger ensembles yield better expected scores. An ensemble-adjusted IS is proposed, which extrapolates the score of an m-member ensemble to the score that the ensemble would achieve if it had fewer or more than m members. Using the ensemble adjustment, a fair version of the IS is derived, which is optimized if ensembles are statistically consistent with the observations. The benefit of the ensemble adjustment is illustrated by comparing ISs of ensembles of different sizes in a seasonal climate forecasting context and a medium-range weather forecasting context. An ensemble-adjusted score can be used for a fair comparison between ensembles of different sizes, and to accurately estimate the expected score of a large operational ensemble by running a much smaller hindcast ensemble.

KEYWORDS
ensembles, scoring rule, seasonal prediction, statistical methods, verification

1 | INTRODUCTION

Weather and climate services routinely issue their forecasts as ensemble forecasts, i.e. collections of forecasts that refer to the same target, but which differ in their initial conditions, boundary conditions, or model formulation (Sivillo *et al.*, 1997). Ensembles can serve as the basis to derive different forecast products, such as point forecasts using the ensemble mean, or probability forecasts using the ensemble mean and standard deviation to forecast a normal distribution (Zhu, 2005). These different forecast products derived from ensembles require different methods of forecast verification (Jolliffe and Stephenson, 2012, chapter 8). In this paper we study the application of probabilistic scoring rules to ensemble forecasts (Winkler *et al.*, 1996; Gneiting and Raftery, 2007).

The Ignorance Score (IS; Roulston and Smith, 2002), also called the Logarithmic Score (Good, 1952; Gneiting and Raftery, 2007), is a strictly proper verification score for probability forecasts. If the forecast is issued as a probability density function $p(z)$, and the forecast target materializes as the value x, then the IS is given by the negative logarithm of the forecast density evaluated at x:

$$\mathcal{I}(p;x) = -\log p(x). \tag{1}$$

The Ignorance difference between two forecasts $\Delta = -\log q(x) + \log p(x)$ implies that the forecast p assigns e^{Δ} times as much density as the forecast q to the observation x. In the negative-log representation of Equation 1, the IS acts as a penalty which a forecaster will try to minimize. Lower scores therefore indicate "better" forecasts. The unit in which Ignorance is measured depends on the base of the logarithm used to calculate the score: *nats* for the natural logarithm, *bits* for base 2 and *bans* for base 10 (MacKay, 2003, section 18.3). The IS has been used as a verification measure for probabilistic forecasts of weather and climate (Barnston *et al.*, 2010; Krakauer *et al.*, 2013; Smith *et al.*, 2015; Rodrigues *et al.*, 2014), and for parameter estimation in dynamical systems (Du and Smith, 2012). The IS has an information-theoretic interpretation (Roulston and Smith, 2002; Peirolo, 2011), and an interpretation in terms of betting returns (Hagedorn and Smith, 2009). Benedetti (2010) shows that "the logarithmic score is the only [verification score] to

respect three basic desiderata [additivity, locality, strict propriety] whose violation can hardly be accepted", and argues that the IS is therefore the "univocal measure of forecast goodness". Due to the locality property, the IS is insensitive to the distance of the verifying observation from the bulk of the forecast distribution. A scoring rule which exhibits sensitivity to distance, such as the Continuous Ranked Probability Score (CRPS; Matheson and Winkler, 1976) might be preferred in certain settings. Furthermore, the IS is relatively insensitive to over-dispersed forecast distributions compared to other scores (Christensen *et al.*, 2014).

If the forecast density is issued as a normal distribution with mean μ and variance σ^2, then the Ignorance is given by

$$\mathcal{I}\left(\mu, \sigma^2; x\right) = \frac{1}{2}\log 2\pi + \frac{1}{2}\log\sigma^2 + \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2, \qquad (2)$$

which follows from the distribution law of the normal distribution (Gneiting *et al.*, 2005). The IS depends on the spread σ of the forecast distribution and on the squared normalized prediction error $[(x-\mu)/\sigma]^2$ of the forecast mean. The score Equation 2 is from the class of proper scoring rules that depend only on the first two moments of the forecast distribution (Dawid and Sebastiani, 1999; Gneiting and Raftery, 2007). Equation 2 thus also applies to non-normal forecast distributions with finite first and second moments μ and σ^2.

Probability forecasts are often generated by running an ensemble of m simulations of a deterministic model to approximate a forecast distribution (Gneiting and Raftery, 2005). There are different possibilities to transform a finite ensemble into a continuous forecast distribution (e.g. Déqué *et al.*, 1994; Gneiting *et al.*, 2005; Bröcker and Smith, 2008). One simple possibility is to transform the ensemble forecast with members $\{y_1, \ldots, y_m\}$ into a normal forecast distribution, whose mean and variance are given by the unbiased estimators of the ensemble mean

$$\widehat{\mu}_m = \frac{1}{m}\sum_{i=1}^{m} y_i \qquad (3)$$

and the ensemble variance

$$\widehat{\sigma}_m^2 = \frac{1}{m-1}\sum_{i=1}^{m}(y_i - \widehat{\mu}_m)^2, \qquad (4)$$

respectively. Mean and variance can be calculated either from the raw ensemble generated by the numerical model directly, or after post-processing the ensemble to correct for systematic forecast errors (such as mean bias or error in the trend). Transformation to a normal distribution only takes into account the first and second moment of the ensemble; any higher-order forecast information (such as skewness or multi-modality) is ignored.

Suppose a forecaster chooses to transform an m-member ensemble forecast to a normal forecast distribution with mean $\widehat{\mu}_m$ and variance $\widehat{\sigma}_m^2$. If the forecast target materializes as the value x, the IS of this forecast is

$$\mathcal{I}(\widehat{\mu}_m, \widehat{\sigma}_m^2; x) = \frac{1}{2}\log 2\pi + \frac{1}{2}\log\widehat{\sigma}_m^2 + \frac{(x-\widehat{\mu}_m)^2}{2\widehat{\sigma}_m^2}. \qquad (5)$$

The score Equation 5 can be interpreted as an evaluation of the forecast distribution $\mathcal{N}(\widehat{\mu}_m, \widehat{\sigma}_m^2)$, or as an evaluation of the underlying m-member ensemble based on its first two sample moments. Note that, since the ensemble members $\{y_1, \ldots, y_m\}$ are assumed to be random variables, the sample mean $\widehat{\mu}_m$, the sample variance $\widehat{\sigma}_m^2$, and the IS given by Equation 5 are random variables too.

In section 2 we will show that, even though the estimators $\widehat{\mu}_m$ and $\widehat{\sigma}_m^2$ are unbiased, i.e. $\mathbb{E}(\widehat{\mu}_m) = \mu$ and $\mathbb{E}(\widehat{\sigma}_m^2) = \sigma^2$, the expected value of $\mathcal{I}(\widehat{\mu}_m, \widehat{\sigma}_m^2; x)$ is not equal to $\mathcal{I}(\mu, \sigma^2; x)$. That is, the IS estimated for a finite ensemble by Equation 5 is, on average, different from the IS that the corresponding normal distribution $\mathcal{N}(\mu, \sigma^2)$ would achieve for the same verifying observation x. Furthermore, the average difference between $\mathcal{I}(\widehat{\mu}_m, \widehat{\sigma}_m^2; x)$ and $\mathcal{I}(\mu, \sigma^2; x)$ depends on the ensemble size. These features are problematic, since they make it difficult to compare the quality of ensembles with different ensemble sizes, and they imply that the IS favours ensembles that are not statistically consistent with the observations.

The following example illustrates the finite ensemble behaviour of the IS. Suppose verifying observations x have a standard normal distribution $\mathcal{N}(0, 1)$. If the standard normal distribution is used as a forecast distribution for x, the expected IS equals

$$\mathbb{E}\left[\mathcal{I}(0, 1; x)\right] = \frac{1}{2}(\log 2\pi + 1) \approx 1.42. \qquad (6)$$

Now suppose m ensemble members are drawn independently from the standard normal distribution. A normal forecast distribution $\mathcal{N}(\widehat{\mu}_m, \widehat{\sigma}_m^2)$ is derived from the ensemble, with mean and variance estimated by Equations 3 and 4, and we calculate the expected IS of this forecast. Note that ensemble members and observations are both random quantities in this setting. The expected IS is thus calculated by taking expectation first over the ensemble members at a given value of x, and then taking expection over the observations x. Alternatively, the expected score can be approximated by averaging over many randomly drawn realizations of forecasts and observations.

In Figure 1 the finite ensemble effect of the Ignorance is illustrated for this forecast setting. The expected IS is shown as a function of the ensemble size m, and compared to the expected IS achieved by the standard normal distribution (the "true" forecast distribution that generated the ensemble members). We have approximated the expectation by simulating 10^5 ensemble–observation pairs for a few values of m. We have also calculated the expectation analytically, using the results presented in section 2. Figure 1 shows that the expected IS of the forecast distribution $\mathcal{N}(\widehat{\mu}_m, \widehat{\sigma}_m^2)$ differs systematically from the expected IS of the distribution $\mathcal{N}(0, 1)$ from which the ensemble members were drawn. The difference in the expected score is especially large for small ensembles. For five-member ensembles, the scores differ by more than 0.5 nats, that is, the standard normal distribution assigns on average $\exp(0.5) \approx 1.65$ times as much probability to the verifying observation than the normal distribution estimated

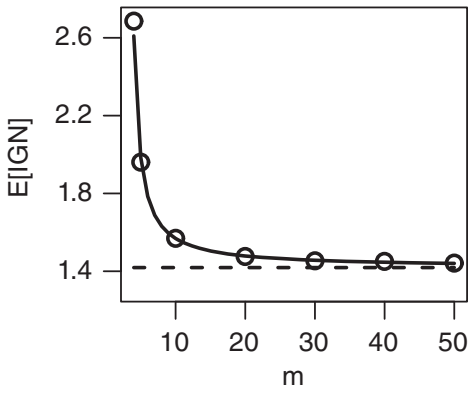

FIGURE 1 Observations are assumed to have a standard normal distribution. The dashed line depicts the expected value of the Ignorance Score (IS) if the standard normal distribution is used as a forecast. The circles depict the average of the IS over 10^5 samples for ensemble sizes $m = 4, 5, 10, 20, \ldots, 50$, calculated by Equation 5, where ensembles are drawn from the standard normal distribution. The solid line depicts the mathematical expectation of the IS as a function of the ensemble size m

from a five-member ensemble whose members have distribution $\mathcal{N}(0, 1)$.

The finite ensemble effect of the IS, its adjustment, and its practical implications for ensemble verification are the main subjects of this paper. The impact of ensemble size on forecast performance was studied for example by Buizza and Palmer (1998), who found that increasing the ensemble size improves various verification measures. The effect of ensemble size on probabilistic verification measures, as well as adjustments for the finite ensemble effect were studied in more detail, for example, by Ferro (2007) for the Brier Score, by Ferro *et al.* (2008) for the discrete and continuous ranked probability score, by Müller *et al.* (2005) for the ranked probability skill score, and by Richardson (2001) for the reliability diagram, the Brier (Skill) score and potential economic value. Further discussions of finite-sample effects on verification scores for ensemble forecasts can be found in Fricker *et al.* (2013) and Ferro (2014).

In section 2, the ensemble-adjusted IS for normal distributions is derived. The score accounts for the finite-ensemble effect by adjusting the IS of an m-member ensemble to the score that would be achieved if the ensemble had fewer or more members. In section 3 the fair IS is derived, which estimates the IS of a hypothetical infinitely large ensemble. The score is fair, because it favours ensembles that are statistically consistent with the observations, which is not the case for the unadjusted score. In section 4.1, the ensemble-adjusted IS is applied to seasonal ensemble hindcasts of European summer temperatures. It is shown that a 41-member ensemble yields a better unadjusted IS than a 10-member ensemble, and that the score of the 10-member ensemble can be adjusted to correctly estimate the expected score of the 41-member ensemble. Section 4.2 strengthens this finding by application to a much larger dataset of medium-range weather forecasts, which is also used to discuss the effect of non-normality on the accuracy of the score adjustment. Section 5 concludes with a discussion of further possible applications of the ensemble-adjusted IS.

2 | THE ENSEMBLE-ADJUSTED IGNORANCE SCORE

In order to account for the finite ensemble effect of the IS, we have to make statistical assumptions about the ensemble. In previous work on ensemble verification (e.g. Anderson, 1996, Hamill, 2001, Siegert *et al.*, 2012, Ferro, 2014), it has been argued that ensemble members should be interpreted as independent draws from a hypothetical "underlying distribution". This interpretation assumes that there is a (possibly infinitely large) population of possible ensemble members, and the realized ensemble is an independent random sample drawn from this population. This picture of ensemble forecasts captures the inherent, unpredictable variability of the ensemble due to the chaoticity of the simulated system. Furthermore, the concept of statistical exchangeability is captured, meaning that the ensemble members are statistically indistinguishable from each other. The hypothetical underlying distribution that generated an ensemble is generally distinct from any forecast distribution that was derived from the realized ensemble.

In this study, it will be assumed that the ensemble members are independent identically distributed (i.i.d.) samples from a normal distribution $\mathcal{N}(\mu, \sigma^2)$ with unknown mean and variance parameters. Normality of the ensemble seems like an overly strong assumption. We might expect ensembles generated by complex system simulations, to have skewed distributions, or heavier tails than a normal distribution, or be multi-modal. But the normality assumption is already inherent when a normal distribution is fitted to the ensemble to issue the probabilistic forecast $\mathcal{N}(\widehat{\mu}_m, \widehat{\sigma}_m^2)$. If there were strong evidence against normality of the ensemble, a normal forecast distribution should not be issued, and the IS would not be calculated according to Equation 2. The results presented in this study are therefore restricted to cases where the ensemble can be reasonably assumed to be normally distributed, such that the normal forecast distribution is a reasonable choice. In highly nonlinear forecast situations, an initial normal distribution is quickly distorted into a non-normal distribution, and so the theory developed here does not strictly apply. In section 4.2 we therefore study deviations from normality in a realistic forecast setting, as well as the effects of such deviations on the proposed score adjustment. The assumption that ensemble members are i.i.d. draws from a distribution is justified if the ensemble has been initialized from randomly drawn samples from some error distribution. But the independence assumption might be violated in ensembles that have been post-processed to correct systematic model biases. While subtracting a constant from each ensemble member leaves the independence assumption intact, a general affine transformation (Bröcker and Smith,

2008) is likely to introduce dependencies between ensemble members.

For the rest of the paper we will refer to $\mathcal{I}(\mu, \sigma^2; x)$ as the *population Ignorance Score*. The population IS the score that the underlying distribution $\mathcal{N}(\mu, \sigma^2)$, or equivalently, a normal distribution $\mathcal{N}(\hat{\mu}_m, \hat{\sigma}_m^2)$ estimated from an infinitely large ensemble would achieve. We will further refer to the score $\mathcal{I}(\hat{\mu}_m, \hat{\sigma}_m^2; x)$, which evaluates the normal distribution derived from a finite m-member ensemble, as the *unadjusted Ignorance Score*. The finite-ensemble effect on the unadjusted IS will be calculated explicitly in this section, and the *ensemble-adjusted Ignorance Score* $\mathcal{I}_{m\to M}$ is derived. The ensemble-adjusted IS allows us to use an m-member ensemble to estimate the expected score that an M-member ensemble drawn from the same underlying distribution would achieve.

We first recall standard results from parameter estimation in normal distributions. Under the assumption that the ensemble members $\{y_1, \ldots, y_m\}$ are i.i.d. draws from a normal distribution $\mathcal{N}(\mu, \sigma^2)$, the sampling distributions of $\hat{\mu}_m$ and $\hat{\sigma}_m^2$, as calculated by Equations 3 and 4, are given by

$$\hat{\mu}_m \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{m}\right) \tag{7}$$

and

$$\frac{m-1}{\sigma^2}\hat{\sigma}_m^2 \sim \chi^2_{m-1}, \tag{8}$$

where χ^2_{m-1} denotes the χ^2-distribution with $m-1$ degrees of freedom. Furthermore, $\hat{\mu}_m$ and $\hat{\sigma}_m^2$ are statistically independent. For proofs refer to Mood (1950, section 4.3)

To calculate the finite-ensemble effect of the unadjusted IS, we calculate the expectations of $\log \hat{\sigma}_m^2$ and $(\hat{\mu}_m - x)^2/\hat{\sigma}_m^2$ using the sampling distributions of $\hat{\mu}_m$ and $\hat{\sigma}_m^2$. In Appendices A.1 and A.2 it is shown that the expectations are

$$\mathbb{E}\left[\log \hat{\sigma}_m^2\right] = \log \sigma^2 + \Psi\left(\frac{m-1}{2}\right) - \log\left(\frac{m-1}{2}\right), \tag{9}$$

and

$$\mathbb{E}\left[\frac{(\hat{\mu}_m - x)^2}{\hat{\sigma}_m^2}\right] = \frac{m-1}{m-3}\left(\frac{\mu - x}{\sigma}\right)^2 + \frac{m-1}{m(m-3)}, \tag{10}$$

where $\Psi(x)$ is the digamma function[1]. Note that Equation 10 only holds for $m \geq 4$; otherwise the expectation is undefined due to the diverging second-moment of the t-distribution (cf. Appendix A.2). If we assume that ensemble members are statistically equivalent to i.i.d. draws from a normal distribution, Equation 10 implies that the unadjusted IS evaluated for ensembles with less than four members has infinite expectation.

Using Equations 9 and 10, the expectations of $\log \hat{\sigma}_M^2$ and $(\hat{\mu}_M - x)^2(\hat{\sigma}_M^2)^{-1}$ can respectively be written in terms of the expectations of $\log \hat{\sigma}_m^2$ and $(\hat{\mu}_m - x)^2(\hat{\sigma}_m^2)^{-1}$, i.e. where $\hat{\mu}_m$ and $\hat{\sigma}_m^2$ were calculated from a different ensemble size $m \neq M$.

[1]Numerical approximations of the digamma function are widely implemented in scientific software, for example `digamma(x)` in R, and `special.psi(x)` in SciPy.

With these results, the ensemble-adjusted IS can be derived, given by

$$\begin{aligned}\mathcal{I}_{m\to M}(\hat{\mu}_m, \hat{\sigma}_m^2; x) = {} & \frac{1}{2}\log 2\pi + \frac{1}{2}\log \hat{\sigma}_m^2 \\ & + \frac{1}{2}\left(\frac{M-1}{M-3}\right)\left(\frac{m-3}{m-1}\right)\frac{(\hat{\mu}_m - x)^2}{\hat{\sigma}_m^2} + \frac{(m-M)(M-1)}{2Mm(M-3)} \\ & + \frac{1}{2}\left[\Psi\left(\frac{M-1}{2}\right) - \Psi\left(\frac{m-1}{2}\right) + \log\left(\frac{m-1}{M-1}\right)\right].\end{aligned} \tag{11}$$

The ensemble-adjusted IS depends on the mean and variance estimated from an m-member ensemble. But the score is adjusted to have expectation equal to the expected IS achieved by the normal distribution $\mathcal{N}(\hat{\mu}_M, \hat{\sigma}_M^2)$ whose mean and variance were estimated from an M-member ensemble, i.e.

$$\mathbb{E}\left[\mathcal{I}_{m\to M}(\hat{\mu}_m, \hat{\sigma}_m^2; x)\right] = \mathbb{E}\left[\mathcal{I}(\hat{\mu}_M, \hat{\sigma}_M^2; x)\right]. \tag{12}$$

Equation 12 can be verified using Equations 9 and 10. The ensemble-adjusted IS is an unbiased estimator of the score that the ensemble would achieve if it had fewer or more than m members. Note that for $M = m$, the unadjusted IS $\mathcal{I}(\hat{\mu}_m, \hat{\sigma}_m^2; x)$ is recovered.

3 | THE FAIR IGNORANCE SCORE

Fricker *et al.* (2013) and Ferro (2014) have introduced the concept of fair verification scores for ensemble forecasts. A fair verification score is optimized if the members of the evaluated ensemble behave like draws from the same distribution as the verifying observation. In other words, fair scores are optimized if the ensemble members are statistically consistent with the observations. It should be emphasized that "fairness" is a property of verification scores for ensemble forecasts, not probability forecasts (although the two are closely related).

By taking the limit $M \to \infty$ in Equation 11 (and using $\lim_{x\to\infty}[\Psi(x) - \log(x)] = 0$), the *fair Ignorance Score* is derived, given by

$$\begin{aligned}\mathcal{I}_{m\to\infty}(\hat{\mu}_m, \hat{\sigma}_m^2; x) = {} & \frac{1}{2}\log 2\pi + \frac{1}{2}\log \hat{\sigma}_m^2 + \frac{1}{2}\left(\frac{m-3}{m-1}\right)\frac{(\hat{\mu}_m - x)^2}{\hat{\sigma}_m^2} \\ & - \frac{1}{2}\left[\Psi\left(\frac{m-1}{2}\right) - \log\left(\frac{m-1}{2}\right) + \frac{1}{m}\right].\end{aligned} \tag{13}$$

The fair IS is an unbiased estimator of the population IS, i.e.

$$\mathbb{E}[\mathcal{I}_{m\to\infty}(\hat{\mu}_m, \hat{\sigma}_m^2; x)] = \mathcal{I}(\mu, \sigma^2; x), \tag{14}$$

which can be verified using Equations 9 and 10. The IS $\mathcal{I}(\mu, \sigma^2; x)$ is a strictly proper verification score for probability forecasts; it therefore holds that the expectation $\mathbb{E}\left[\mathcal{I}(\mu, \sigma^2; x)\right]$ is minimized if and only if $x \sim \mathcal{N}(\mu, \sigma^2)$, that is, if the forecast distribution is equal to the distribution of the observation (Gneiting and Raftery, 2007). Recall that μ and σ^2 are the parameters of the hypothetical normal distribution from which the ensemble members $y_1, ..., y_m$ were independently drawn. It thus follows that the expected fair IS $\mathbb{E}\left[\mathcal{I}_{m\to\infty}(\hat{\mu}_m, \hat{\sigma}_m^2; x)\right]$ is optimized if and only if the ensemble members and the observation behave like draws from the same (normal) distribution.

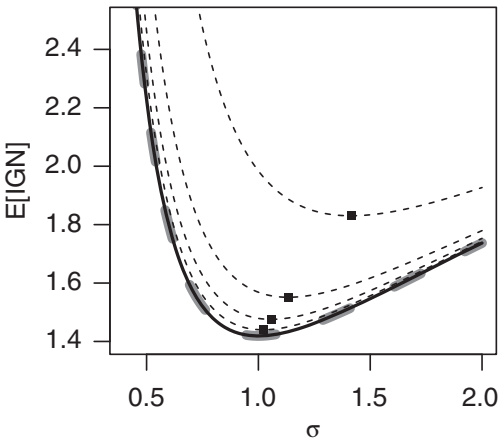

FIGURE 2 Verifications x are drawn i.i.d. from $\mathcal{N}(0,1)$, and m-member ensembles are drawn i.i.d. from $\mathcal{N}(0,\sigma^2)$. The dashed grey line corresponds to the expected IS for $m \to \infty$, the black dashed lines correspond to the expected value of the unadjusted ISs of m-member ensembles (from top to bottom: $m = 5, 10, 20, 50$), and the black full line shows the expected fair IS (independent of m). Square markers indicate the minima along their respective curves

Figure 2 illustrates differences between the expected values of the unadjusted and the fair IS when ensembles and observations are not drawn from the same distributions. The observation is assumed to have a standard normal distribution $\mathcal{N}(0,1)$, and the m-member ensemble is assumed to be drawn from the ensemble distribution $\mathcal{N}(0,\sigma^2)$, i.e. the ensembles are not statistically consistent with the observations, except when $\sigma^2 = 1$. The expectations of the unadjusted IS (for different ensemble sizes) of the population IS and of the fair IS are shown as functions of the standard deviation σ of the ensemble distribution. The expectations were calculated analytically using Equations 9 and 10, but they can also be approximated by averaging scores calculated for large numbers of simulated forecasts and observations. The systematic effect due to the finiteness of the ensemble shows as a vertical offset of the curves. The expected value of the unadjusted IS is larger when the ensemble size is smaller. At any given value of σ, the expectation of the unadjusted IS can be improved by generating a larger ensemble. Conversely, the expected fair IS is independent of the ensemble size, and equals the expected population IS for all values of m and σ.

Figure 2 further shows that the unadjusted IS is not a fair verification score. The unadjusted IS obtains its optimum at a value of σ which differs from the standard deviation of the distribution $\mathcal{N}(0,1)$ of the observation. The unadjusted IS thus rewards ensembles that violate statistical consistency, i.e. whose members do not behave like draws from the same distribution as the observation (Anderson, 1996). The ensemble that optimizes the unadjusted IS is overdispersive, i.e. the ensemble spread overestimates the variability of the observation. Such an ensemble would not pass the rank histogram test for statistical consistency proposed by Anderson (1996); the rank histogram would appear ∩-shaped. On the other hand, the ensemble that optimizes the fair IS will have a flat rank histogram, because the ensemble members are sampled from the same distribution as the observation.

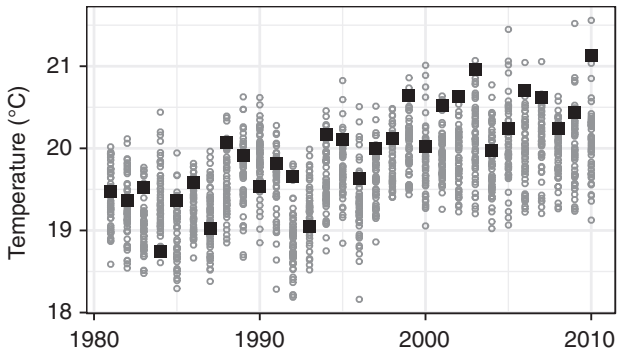

FIGURE 3 Time series of System4 ensemble forecasts (small grey markers) and observations (large black markers) for summer (JJA) surface temperatures averaged over Europe

4 | APPLICATIONS

4.1 | Seasonal predictions of European mean temperature

We illustrate the ensemble-adjusted IS by application to a dataset of retrospective seasonal climate forecasts. We consider ensemble predictions of the summer (JJA) mean air surface temperature over land over the area 30°N–75°N, 12.5°W–42.5°E (roughly Europe), initialized on 01 May of the same year. The forecasts were generated by ECMWF's seasonal forecast system "System4" (Molteni *et al.*, 2011) with start dates from 1981 to 2010 ($n = 30$), and $m = 51$ ensemble members. Verifying observations are taken from the WFDEI gridded dataset (Weedon *et al.*, 2011; Dee *et al.*, 2011). All data were downloaded through the ECOMS user data gateway (ECOMS, 2014). The ensemble and observation data are plotted over time in Figure 3. Visual inspection shows that a normal approximation of the ensemble forecasts is reasonable. The normal assumption is further justified by the approximately uniform distribution of the p-values of Shapiro–Wilk normality tests applied to the individual ensembles (not shown). The System4 ensemble has a slight cold bias of ≈ -0.3 K. The observations show a linear trend of ≈ 0.05 K/year which is underestimated by the trend of the ensemble mean (≈ 0.03 K/year). After removing individual linear trends from observations and ensemble means, the Pearson correlation coefficient between ensemble means and observations is 0.46.

Consider the following scenario where an ensemble-adjusted verification score is useful. A climate centre introduces a new forecast system that can routinely produce operational ensemble forecasts with 41 members. In order to inform forecast users about the quality of the forecast system, a dataset of retrospective forecasts is produced. But computational resources are limited, so that the hindcast dataset can be produced with only 10 ensemble members. The IS averaged

TABLE 1 Evaluation of a ten-member ensemble and a 41-member ensemble generated by System4, using the ensemble-adjusted Ignorance Score (mean scores ± standard error of the mean)

m	$\mathcal{I}^{\text{Sys4}}_{m\to 10}$	$\mathcal{I}^{\text{Sys4}}_{m\to 41}$	$\mathcal{I}^{\text{Sys4}}_{m\to \infty}$
10	0.91 (±0.22)	0.74 (±0.18)	0.71 (±0.17)
41	0.94 (±0.18)	0.76 (±0.15)	0.72 (±0.14)

TABLE 2 Evaluation of the 30-member climatological reference ensemble forecast using the ensemble-adjusted Ignorance Score (mean scores ± standard error of the mean)

$\mathcal{I}^{\text{clim}}_{30\to 10}$	$\mathcal{I}^{\text{clim}}_{30\to 30}$	$\mathcal{I}^{\text{clim}}_{30\to 41}$	$\mathcal{I}^{\text{clim}}_{30\to \infty}$
0.97 (±0.13)	0.87 (±0.11)	0.86 (±0.11)	0.83 (±0.10)

over the hindcasts is to be used to evaluate the performance of the ensemble forecasting system. Due to the finite ensemble effect, the average IS achieved by the 10-member ensembles will likely be higher (i.e. worse) than the average score that a 41-member hindcast ensemble could achieve. To provide a more realistic assessment of the expected skill of the operational 41-member forecast ensemble, it is thus important to account for the difference in ensemble size between the hindcast and the forecast. Otherwise hindcast skill will be a too pessimistic estimate of forecast skill.

The scenario outlined above is mimicked by splitting the available 51-member System4 ensemble into two disjoint subensembles of size 10 and size 41. (The first 10 members in the downloaded database are used for the smaller subensemble, and the remaining members for the larger.) The 10-member ensemble acts as the hindcast dataset that was actually generated. The 41-member ensemble acts as the hindcast dataset that should have been generated to make hindcast skill and forecast skill comparable, but could not be realized due to computational limitations. We use these subensembles to address the following questions:

- Is the average (unadjusted) IS calculated for the 10-member hindcast ensemble representative of the score of a 41-member ensemble?
- If not, can the ensemble-adjusted IS be applied to the 10-member ensemble to estimate the score that would be achieved by a 41-member ensemble?
- What are possible pitfalls if the finite ensemble effect is not taken into account?

Table 1 compares ISs $\mathcal{I}^{\text{Sys4}}_{m\to M}$ for the two hindcast ensembles with $m = \{10, 41\}$, adjusted to ensemble sizes $M = \{10, 41, \infty\}$. We find that if no ensemble-adjustment is applied, i.e. if $m = M$, the larger ensemble with $m = 41$ members achieves a lower average score than the ensemble with $m = 10$ members. The average difference between the ISs of the large and small ensemble is 0.15 (±0.13) nats, indicating that the normal forecast derived from the 41-member ensemble assigns on average 1.16 times more probability to the observation than the normal forecast derived from the 10-member ensemble. Comparing the scores after applying an ensemble adjustment, i.e. comparing $\mathcal{I}^{\text{Sys4}}_{10\to 41}$ and $\mathcal{I}^{\text{Sys4}}_{41\to 41}$, the difference is much smaller at $-0.02(\pm 0.1)$ nats, which indicates that the ensembles would be equally skilful if they had the same number of 41 members. The same conclusion is drawn when comparing $\mathcal{I}^{\text{Sys4}}_{10\to 10}$ with $\mathcal{I}^{\text{Sys4}}_{41\to 10}$, and also when comparing the fair ISs of the ensembles. The two ensembles appear equally skilful after adjusting their scores for the different ensemble size. Since the two ensembles were generated by the same forecasting system, this finding is intuitively reasonable.

A forecast user might be interested in how the System4 ensemble compares to the score of a simple benchmark forecast such as climatology. In an ensemble forecasting context, a climatological forecast can be generated by treating the available 30 observations as a constant 30-member ensemble forecast which is issued every year. In Table 2, the 30-member climatological ensemble forecast is evaluated using the IS adjusted for different ensemble sizes M. We find that forecasts derived from the 10-member System4 ensemble obtain a slightly higher (worse) IS than the forecast derived from the climatological ensemble (i.e. $\mathcal{I}^{\text{Sys4}}_{10\to 10} > \mathcal{I}^{\text{clim}}_{30\to 30}$). Conversely, forecasts derived from the 41-member ensemble obtain a lower (better) IS than climatology (i.e. $\mathcal{I}^{\text{Sys4}}_{41\to 41} < \mathcal{I}^{\text{clim}}_{30\to 30}$). The difference between the 41-member System4 ensemble and climatology is correctly estimated when the IS of the 10-member ensemble is adjusted to 41 members ($\mathcal{I}^{\text{Sys4}}_{10\to 41} < \mathcal{I}^{\text{clim}}_{30\to 30}$). If we compare the unadjusted average scores of the 10-member System4 ensemble and climatology, we conclude that System4 is less skilful than climatology. But this is due to the small hindcast size. By adjusting the score of the 10-member hindcast to the size of the actual forecast ensemble, we correctly conclude that the 41-member System4 ensemble is preferable to climatology. The same conclusion is drawn when the fair ISs are compared, suggesting that the hypothetical underlying distribution of System4 assigns on average more probability to the observation than the climatological distribution.

4.2 | Application to medium-range forecasts

Section 4.1 illustrated the potential practical benefits of using an ensemble-adjusted verification score. But the standard errors are too large to draw definite conclusions about the validity of the adjustment in practice, and about the possible effect of non-normality of the ensembles. The unadjusted and fair ISs have therefore been applied to evaluate a much larger dataset of ensemble forecasts from a numerical experiment with ECMWF's IFS model using 200 members (Leutbecher, 2018). The experiment uses the forecast model version that was operational at ECMWF from March to November 2016. The representation of uncertainties in this research ensemble is the same as that of the operational ensemble at the

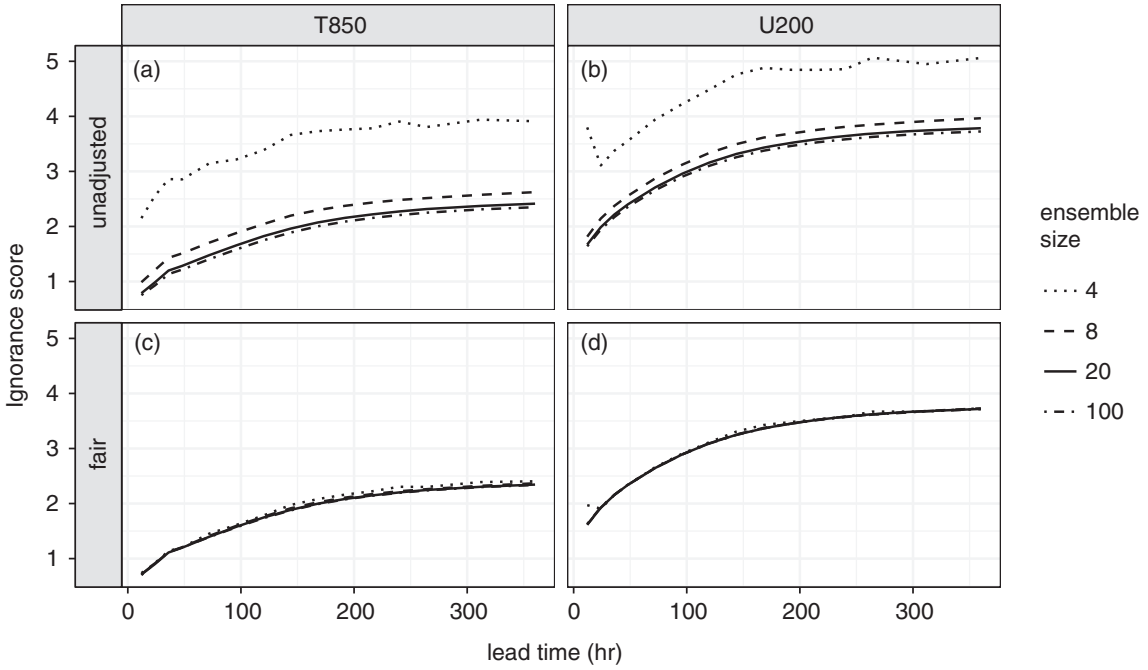

FIGURE 4 (a, b) Unadjusted and (c, d) fair Ignorance Scores for (a, c) *T*850 and (b, d) *U*200, plotted over lead time for different ensemble sizes

time. Initial uncertainties are represented with singular vector perturbations and with perturbations from an ensemble of 4D-Var analyses (Leutbecher and Palmer, 2008; Buizza *et al.*, 2008). Model uncertainties are represented through stochastic perturbations of the model tendencies using two schemes: Stochastically Perturbed Parametrization Tendencies (SPPT) and Stochastic Kinetic Energy Backscatter (SKEB). Further details on the configurations of these schemes are described by Leutbecher *et al.* (2017). Apart from the plus–minus symmetry of the initial perturbations and the ocean initial conditions, the members of this ensemble can be considered independent realizations from the same distribution. This research ensemble uses a horizontal resolution of 29 km and contains 200 members while the operational ensemble has 50 members but a higher resolution of 18 km. The experiment was run for the boreal summer season of June–August 2016 with one forecast issued at 0000 UTC daily.

ISs have been computed for various ensemble sizes from 4 to 200 members. Results are presented for 4, 8, 20 and 100 members for 200 hPa zonal wind (*U*200) and 850 hPa temperature (*T*850) in the Northern Hemisphere extratropics (20°N–90°N). To compute the scores, the forecast fields of upper-air variables have been spectrally truncated to a horizontal wavenumber of 120 prior to transforming the spectral fields to a regular 1.5° latitude–longitude grid. Cosine-latitude weights were used in the spatial averaging of the scores. The ensemble forecasts were verified with operational analysis fields which were transformed to the same regular grid after truncation to wavenumber 120. For each variable and forecast lead time, the total number of forecast–observation pairs is 1.0×10^6 (the effective sample size is smaller due to spatial correlations). For ensemble sizes up to 20 members, only the odd members were selected from the 200-member experiment, in order to obtain ensemble members with i.i.d. atmospheric initial conditions.

Figures 4a,b show unadjusted ISs plotted against lead time for different ensemble sizes, averaged over all grid points and forecast start dates. As expected, smaller ensembles achieve higher ISs than larger ensembles generated by the same forecasting system. The score difference is stable over lead time. Figure 4c,d show the same forecast verification using average fair ISs. For both forecast variables, the difference between scores for different ensemble sizes reduces considerably due to the correction for the systematic effect of ensemble size. It can be noted that similar results were obtained for 850 hPa zonal wind, 500 hPa geopotential, and also for the Tropics and Southern Hemisphere extratropics (not shown).

A key assumption in the derivation of the adjusted IS was that the forecast ensembles are draws from a normal distribution, which is not necessarily the case in forecasts produced by a numerical model of atmospheric dynamics. For non-Gaussian ensembles we should not expect the adjustment to be exact. The good agreement between fair scores for different ensemble sizes might be explained by the ensemble forecasts being practically indistinguishable from normal random variables. But applying a Shapiro–Wilk Normality test to the 200 member ensembles revealed that 38% of *U*200 ensemble forecasts have a *p*-value less than 0.05 (64% for *T*850), which suggests a considerable number of significantly non-Gaussian ensembles. We therefore hypothesize that the ensemble adjustment must also be valid for ensembles that are "not too far away" from normality.

To test this hypothesis, we further calculated skewness and kurtosis of the 200-member ensemble forecasts. For each ensemble, we also calculated the difference between the fair score of the 200-member ensemble and the fair score of the

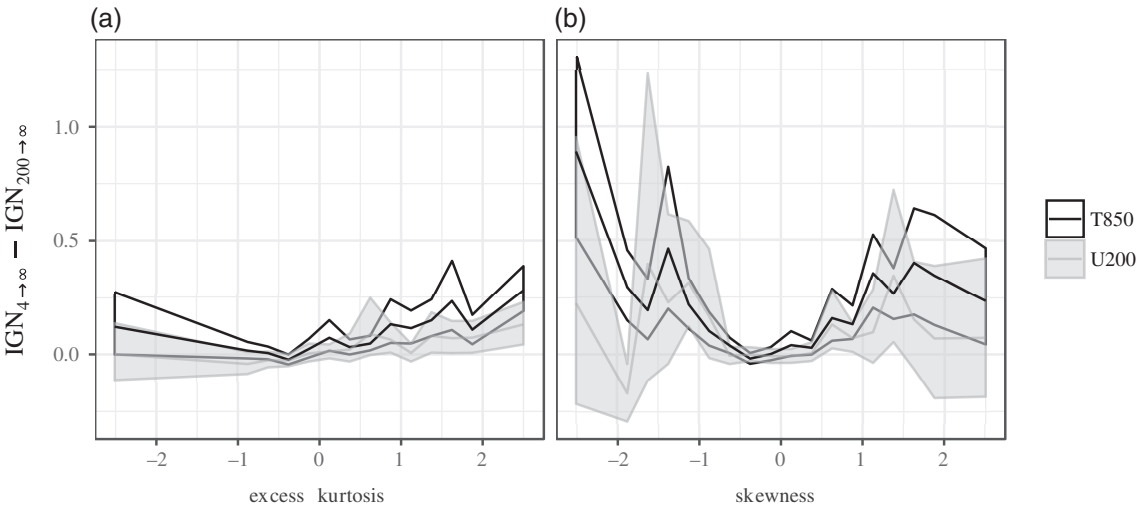

FIGURE 5 Averages and 95% confidence bands of the Ignorance difference $\mathcal{I}_{4\to\infty} - \mathcal{I}_{200\to\infty}$ (measured in nats) stratified by (a) kurtosis and (b) skewness for $T850$ and $U200$ forecasts, at lead time 5 days. The mean score difference vanishes for ensembles with zero kurtosis or zero skewness. For non-normal ensembles, the four-member ensemble tends to obtain higher scores than the 200-member ensemble. To interpret the magnitude of these Ignorance differences, note that the overall average difference between the unadjusted scores $\mathcal{I}(\widehat{\mu}_4, \widehat{\sigma}_4^2; x) - \mathcal{I}(\widehat{\mu}_{200}, \widehat{\sigma}_{200}^2; x)$ is 1.75 nats for $T850$ and 1.25 nats for $U200$. The sampling distribution of kurtosis is such that 90% of all kurtosis values are in $[-0.60, 1.54]$ for $U200$ and in $[-0.81, 2.96]$ for $T850$ (respectively $[-0.56, 0.72]$ and $[-1.01, 1.02]$ for skewness). Deviations from zero kurtosis tend to coincide with deviations from zero skewness; the correlation between absolute skewness and absolute kurtosis is 0.62 for $U200$ forecasts and 0.70 for $T850$ forecasts.

subsampled four-member ensemble,

$$\mathcal{I}_{4\to\infty}(\widehat{\mu}_4, \widehat{\sigma}_4^2; x) - \mathcal{I}_{200\to\infty}(\widehat{\mu}_{200}, \widehat{\sigma}_{200}^2; x).$$

For perfectly Gaussian ensembles, the average difference between the fair ISs at different ensemble sizes should be zero, but for ensembles that deviate from normality the average difference can be non-zero. In Figure 5 we show the average score difference for the forecast variables $U200$ and $T850$, stratified by

$$\text{skewness} \quad \left(m^{-1} \sum_{i=1}^{m} (y_i - \widehat{\mu}_m)^3 (\widehat{\sigma}_m^2)^{-3/2} \right)$$

$$\text{and excess kurtosis} \quad \left(m^{-1} \sum_{i=1}^{m} (y_i - \widehat{\mu}_m)^4 (\widehat{\sigma}_m^2)^{-2} - 3 \right).$$

95% confidence intervals were estimated from the bootstrap distribution using block-bootstrapping from the 92 start dates without resampling any data in space to preserve spatial correlation. This approach will likely underestimate the spatial degrees of freedom in the data and therefore produce underconfident (too wide) uncertainty intervals. For ensembles with skewness and excess kurtosis close to zero, the average score difference is indeed very close to zero, and increases as skewness and excess kurtosis deviate from zero in either direction. The average mismatch is non-negative for all deviations from normality, i.e. under the fair IS, a non-Gaussian 200-member ensemble scores better, on average, than an equivalent four-member ensemble. This implies that the good match between the average fair ISs of 200-member and four-member ensembles cannot be explained by a cancellation of biases of different signs for, say, positively and negatively skewed ensembles. Furthermore, the score adjustment "fails in the right direction"; a 200-member ensemble obtains a better score than a four-member ensemble from the same distribution, which is what we would expect from the unadjusted score.

5 | SUMMARY AND OUTLOOK

We have studied the application of the well-known IS for ensemble verification. We focused on forecasts issued as normal distributions whose parameters are estimated from the ensemble. It was shown that the IS is sensitive to the number of ensemble members. Forecasts derived from larger ensembles obtain better scores. In section 2 a new estimator of the IS was derived which includes an adjustment for the finite ensemble size. The ensemble-adjusted IS allows us to estimate the IS that an m-member ensemble would achieve if it had fewer or more than m members. In section 3, the special case $M \to \infty$ was argued to yield a fair verification score, which is optimized if ensemble members and observations behave like draws from the same distribution. The benefit of the ensemble-adjustment of the IS was illustrated in section 4.1 by application to seasonal climate forecasts. If the ensemble size of the hindcast is smaller than the ensemble size of the forecast, hindcast skill underestimates forecast skill. By using the ensemble-adjusted IS, the score of a 41-member hindcast could be correctly estimated from a 10-member hindcast ensemble. For the medium-range forecasts analysed in section 4.2, the adjustment of the IS reduces the average score differences from nearly 2 nats to less than 0.1 nats when comparing a four-member ensemble with a 200-member ensemble. The analysis of section 4.2

further showed that the score and its adjustment are robust to (moderate) deviations from normality.

The ensemble-adjusted score has further applications. The difference between the unadjusted IS $\mathcal{I}_{m\to m}$ and the fair IS $\mathcal{I}_{m\to\infty}$ can be interpreted as an information deficit due to the finiteness of the ensemble, which can be estimated without knowing the underlying distribution. Quantifying this information deficit is relevant in information-theoretic predictability frameworks, such as DelSole and Tippett (2007). The hypothetical "underlying distribution" of the ensemble is not available for forecasting, but estimating its forecast skill by the fair IS is of interest to forecast model developers, who might be interested in the performance of the forecasting system independent of ensemble size. If a forecast centre releases a new cycle of their model, with updates to the model physics and ensemble size, the finite-ensemble correction can be used to differentiate how much of the improvement is due to better physics, and how much is due to increased ensemble size. Moreover, it was shown that the unadjusted IS favours unreliable forecast models, i.e. models that produce ensemble members with different statistical properties than the observations. The fair IS is optimized if ensemble members have the same statistical properties as the observations. The fair IS might therefore be a more suitable objective function for tuning parameters of the numerical forecast model, and for estimating parameters in complex systems (e.g. Du and Smith (2012)).

ACKNOWLEDGEMENTS

We are grateful for stimulating discussions with the members of the statistical science group of the University of Exeter. This work was partly funded by the European Union Programme FP7/2007-13 under grant agreement 3038378 (SPECS). Comments from two reviewers and the editor greatly helped to improve the quality of the article.

ORCID

Stefan Siegert http://orcid.org/0000-0001-8938-2823
Christopher A. T. Ferro http://orcid.org/0000-0002-9830-9270
Martin Leutbecher http://orcid.org/0000-0003-4160-0750

REFERENCES

Anderson, J.L. (1996) A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *Journal of Climate*, 9(7), 1518–1530. https://doi.org/10.1175/1520-0442(1996)009<1518:AMFPAE>2.0.CO;2.

Barnston, A.G., Li, S., Mason, S.J., DeWitt, D.G., Goddard, L. and Gong, X. (2010) Verification of the first 11 years of IRI's seasonal climate forecasts. *Journal of Applied Meteorology and Climatology*, 49(3), 493–520. https://doi.org/10.1175/2009jamc2325.1.

Benedetti, R. (2010) Scoring rules for forecast verification. *Monthly Weather Review*, 138(1), 203–211. https://doi.org/10.1175/2009mwr2945.1.

Bröcker, J. and Smith, L.A. (2008) From ensemble forecasts to predictive distribution functions. *Tellus A*, 60(4), 663–678. https://doi.org/10.1111/j.1600-0870.2008.00333.x.

Buizza, R. and Palmer, T.N. (1998) Impact of ensemble size on ensemble prediction. *Monthly Weather Review*, 126(9), 2503–2518. https://doi.org/10.1175/1520-0493(1998)126<2503:IOESOE>2.0.CO;2.

Buizza, R., Leutbecher, M. and Isaksen, L. (2008) Potential use of an ensemble of analyses in the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 134, 2051–2066.

Christensen, H.M., Moroz, I.M. and Palmer, T.N. (2014) Evaluation of ensemble forecast uncertainty using a new proper score: application to medium-range and seasonal forecasts. *Quarterly Journal of the Royal Meteorological Society*, 141(687), 538–549. https://doi.org/10.1002/qj.2375.

Dawid, A.P. and Sebastiani, P. (1999) Coherent dispersion criteria for optimal experimental design. *Annals of Statistics*, 27(1), 65–81.

Dee, D.P., Uppala, S., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A.C.M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A.J., Haimberger, L., Healy, S.B., Hersbach, H., Hólm, E.V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., Mcnally, A.P., Monge-Sanz, B.M., Morcrette, J.-J., Park, B.K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N. and Vitart, F. (2011) The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137, 553–597. https://doi.org/10.1002/qj.828.

DelSole, T. and Tippett, M.K. (2007) Predictability: recent insights from information theory. *Reviews of Geophysics*, 45(RG4002). https://doi.org/10.1029/2006RG000202.

Déqué, M., Royer, J., Stroe, R. and France, M. (1994) Formulation of Gaussian probability forecasts based on model extended-range integrations. *Tellus A*, 46(1), 52–65. https://doi.org/10.1034/j.1600-0870.1994.00005.x.

Du, H. and Smith, L.A. (2012) Parameter estimation through ignorance. *Physical Review E*, 86(016213). https://doi.org/10.1103/physreve.86.016213.

ECOMS. (2014) *The ECOMS User Data Gateway*. Available at: http://meteo.unican.es/ecoms-udg [Accessed 25th June 2014]

Ferro, C.A.T. (2007) Comparing probabilistic forecasting systems with the Brier score. *Weather and Forecasting*, 22(5), 1076–1088. https://doi.org/10.1175/WAF1034.1.

Ferro, C.A.T. (2014) Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 140, 1917–1923. https://doi.org/10.1002/qj.2270.

Ferro, C.A.T., Richardson, D.S. and Weigel, A.P. (2008) On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorological Applications*, 15(1), 19–24. https://doi.org/10.1002/met.45.

Fricker, T.E., Ferro, C.A.T. and Stephenson, D.B. (2013) Three recommendations for evaluating climate predictions. *Meteorological Applications*, 20(2), 246–255. https://doi.org/10.1002/met.1409.

Gneiting, T. and Raftery, A.E. (2005) Weather forecasting with ensemble methods. *Science*, 310(5746), 248–249. https://doi.org/10.1126/science.1115255.

Gneiting, T. and Raftery, A.E. (2007) Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378. https://doi.org/10.1198/016214506000001437.

Gneiting, T., Raftery, A.E., Westveld, A.H.I. and Goldman, T. (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133(5), 1098–1118. https://doi.org/10.1175/MWR2904.1.

Good, I.J. (1952) Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 14, 107–114. Available at: http://www.jstor.org/stable/2984087 [Accessed 19th December 2018].

Hagedorn, R. and Smith, L.A. (2009) Communicating the value of probabilistic forecasts with weather roulette. *Meteorological Applications*, 16(2), 143–155. https://doi.org/10.1002/met.92.

Hamill, T.M. (2001) Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129(3), 550–560. https://doi.org/10.1175/1520-0493(2001)129<0550:iorhfv>2.0.co;2.

Hogben, D., Pinkham, R. and Wilk, M. (1961) The moments of the non-central *t*-distribution. *Biometrika*, 48(3–4), 465–468. https://doi.org/10.2307/2332772.

Jolliffe, I.T. and Stephenson, D.B. (2012) *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Chichester: John Wiley & Sons.

Krakauer, N.Y., Grossberg, M.D., Gladkova, I. and Aizenman, H. (2013) Information content of seasonal forecasts in a changing climate. *Advances in Meteorology*, 2013, 1–12. https://doi.org/10.1155/2013/480210.

Lehmann, E.L. and Casella, G. (1998) *Theory of Point Estimation*. New York, NY: Springer.

Leutbecher, M. (2018) Ensemble size: how suboptimal is less than infinity?. *Quarterly Journal of the Royal Meteorological Society*, 145(Supplement 1), 107–128.

Leutbecher, M. and Palmer, T.N. (2008) Ensemble forecasting. *Journal of Computational Physics*, 227, 3515–3539. https://doi.org/10.1016/j.jcp.2007.02.014.

Leutbecher, M., Lock, S.J., Ollinaho, P., Lang, S.T.K., Balsamo, G., Bechtold, P., Bonavita, M., Christensen, H.M., Diamantakis, M., Dutra, E., English, S., Fisher, M., Forbes, R.M., Goddard, J., Haiden, T., Hogan, R.J., Juricke, S., Lawrence, H., MacLeod, D., Magnusson, L., Malardel, S., Massart, S., Sandu, I., Smolarkiewicz, P.K., Subramanian, A., Vitart, F., Wedi, N. and Weisheimer, A. (2017) Stochastic representations of model uncertainties at ECMWF: state of the art and future vision. *Quarterly Journal of the Royal Meteorological Society*, 143, 2315–2339. https://doi.org/10.1002/qj.3094.

MacKay, D. (2003) *Information Theory, Inference, and Learning Algorithms*. Cambridge: Cambridge University Press.

Matheson, J.E. and Winkler, R.L. (1976) Scoring rules for continuous probability distributions. *Management Science*, 22(10), 1087–1096. https://doi.org/10.1287/mnsc.22.10.1087.

Molteni, F., Stockdale, T., Balmaseda, M., Balsamo, G., Buizza, R., Ferranti, L., Magnusson, L., Mogensen, K., Palmer, T.N. and Vitart, F. (2011) *The new ECMWF seasonal forecast system (System 4)*. Technical Memorandum No. 656, ECMWF, Reading, UK.

Mood, A.M. (1950) *Introduction to the Theory of Statistics*. New York, NY: McGraw-Hill.

Müller, W., Appenzeller, C., Doblas-Reyes, F. and Liniger, M. (2005) A debiased ranked probability skill score to evaluate probabilistic ensemble forecasts with small ensemble sizes. *Journal of Climate*, 18(10), 1513–1523. https://doi.org/10.1175/JCLI3361.1.

Peirolo, R. (2011) Information gain as a score for probabilistic forecasts. *Meteorological Applications*, 18(1), 9–17. https://doi.org/10.1002/met.188.

Richardson, D.S. (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. https://doi.org/10.1002/qj.49712757715.

Rodrigues, L.R.L., García-Serrano, J. and Doblas-Reyes, F. (2014) Seasonal forecast quality of the West African monsoon rainfall regimes by multiple forecast systems. *Journal of Geophysical Research: Atmospheres*, 119(13), 7908–7930. https://doi.org/10.1002/2013jd021316.

Roulston, M.S. and Smith, L.A. (2002) Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, 130(6), 1653–1660. https://doi.org/10.1175/1520-0493(2002)130<1653:EPFUIT>2.0.CO;2.

Siegert, S., Bröcker, J. and Kantz, H. (2012) Rank histograms of stratified Monte Carlo ensembles. *Monthly Weather Review*, 140(5), 1558–1571. https://doi.org/10.1175/mwr-d-11-00302.1.

Sivillo, J.K., Ahlquist, J.E. and Toth, Z. (1997) An ensemble forecasting primer. *Weather and Forecasting*, 12(4), 809–818. https://doi.org/10.1175/1520-0434(1997)012<0809:AEFP>2.0.CO;2.

Smith, L.A., Du, H., Suckling, E.B. and Niehörster, F. (2015) Probabilistic skill in ensemble seasonal forecasts. *Quarterly Journal of the Royal Meteorological Society*, 141, 1085–1100. https://doi.org/10.1002/qj.2403.

Weedon, G., Gomes, S., Viterbo, P., Shuttleworth, W., Blyth, E., Österle, H., Adam, J., Bellouin, N., Boucher, O. and Best, M. (2011) Creation of the WATCH Forcing data and its use to assess global and regional reference crop evaporation over land during the Twentieth Century. *Journal of Hydrometeorology*, 12, 823–848. https://doi.org/10.1175/2011JHM1369.1.

Winkler, R.L., Munoz, J., Cervera, J.L., Bernardo, J.M., Blattenberger, G., Kadane, J.B., Lindley, D.V., Murphy, A.H., Oliver, R.M. and Ros-Insua, D. (1996) Scoring rules and the evaluation of probabilities. *Test*, 5(1), 1–60. https://doi.org/10.1080/01621459.1969.10501037.

Zhu, Y. (2005) Ensemble forecast: a new approach to uncertainty and predictability. *Advances in Atmospheric Sciences*, 22(6), 781–788. https://doi.org/10.1007/BF02918678.

How to cite this article: Siegert S, Ferro CAT, Stephenson DB, Leutbecher M. The ensemble-adjusted Ignorance Score for forecasts issued as normal distributions. *QJR Meteorol Soc.* 2019;145(Suppl. 1):129–139. https://doi.org/10.1002/qj.3447

APPENDIX

PROOFS

A.1 | $\mathbb{E}\left[\log \hat{\sigma}_m^2\right]$

The derivation follows from the properties of distributions in the exponential family and their sufficient statistics (Lehmann and Casella, 1998, section 1.5). If $X \sim \chi^2_{m-1}$, we can define $\tau := (m-1)/2 - 1$ and write the probability density function of X as

$$p_X(x) = \exp\left\{\tau \log x - \frac{x}{2} - (\tau+1)\log 2 - \log\Gamma(\tau+1)\right\}. \tag{A1}$$

Differentiating the integral $\int \mathrm{d}x\, p_X(x)$ with respect to τ yields

$$\mathbb{E}[\log X] = \log 2 + \Psi\left(\frac{m-1}{2}\right), \tag{A2}$$

where $\Psi(x) = \mathrm{d}/\mathrm{d}x \log\Gamma(x)$ is the digamma function. Applying Equation A2 to $\hat{\sigma}_m^2$, whose distribution is given by Equation 8, we get

$$\mathbb{E}\left[\log \hat{\sigma}_m^2\right] = \mathbb{E}\left[\log \frac{m-1}{\sigma_m^2}\hat{\sigma}^2\right] + \log\frac{\sigma^2}{m-1} \tag{A3}$$

$$= \log\sigma^2 + \Psi\left(\frac{m-1}{2}\right) - \log\left(\frac{m-1}{2}\right). \tag{A4}$$

A.2 | $\mathbb{E}\left[\frac{(\hat{\mu}_m - x)^2}{\hat{\sigma}_m^2}\right]$

Let the independent random variables Z and V have distributions $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi^2_{m-1}$. Then the non-central t-distribution $t_{m-1,x}$, with $m-1$ degrees of freedom and non-centrality parameter x, is defined through

$$\frac{Z + x}{\sqrt{V/(m-1)}} \sim t_{m-1,x}. \tag{A5}$$

Using the sampling distributions of $\hat{\mu}_m$ and $\hat{\sigma}_m^2$, and their independence, we get the following relation:

$$\sqrt{m}\frac{\hat{\mu}_m - x}{\hat{\sigma}_m} = \frac{\dfrac{\hat{\mu}_m - \mu}{\sigma/\sqrt{m}} + \dfrac{\sqrt{m}}{\sigma}(\mu - x)}{\sqrt{\dfrac{m-1}{\sigma^2}\hat{\sigma}_m^2}\Big/\sqrt{m-1}} \tag{A6}$$

$$\sim t_{m-1,\frac{\sqrt{m}}{\sigma}(\mu-x)}. \tag{A7}$$

The raw moments of a random variable $T \sim t_{\nu,z}$ are given by Hogben *et al.* (1961):

$$\mathbb{E}\left[T^k\right] = \left(\frac{\nu}{2}\right)^{\frac{k}{2}} \frac{\Gamma\left(\frac{\nu-k}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \exp\left(-\frac{z^2}{2}\right) \frac{\partial^k}{\partial z^k} \exp\left(\frac{z^2}{2}\right). \quad \text{(A8)}$$

By calculating the second raw moment of $\sqrt{m}(\widehat{\mu}_m - x)/\widehat{\sigma}_m$ and dividing by m, we get

$$\mathbb{E}\left[\frac{(\widehat{\mu}_m - x)^2}{\widehat{\sigma}_m^2}\right] = \frac{m-1}{m-3}\left(\frac{\mu - x}{\sigma}\right)^2 + \frac{m-1}{m(m-3)}. \quad \text{(A9)}$$