Local Proper Scoring Rules

Werner Ehm and Tilmann Gneiting

Institute for Frontier Areas of Psychology and Mental Health
University of Heidelberg and University of Washington

Addendum to Technical Report no. 551

Department of Statistics, University of Washington

February 10, 2010

Philip Dawid, Steffen Lauritzen and Matt Parry have brought to our attention that the Fisher score arises as the special case of the scoring rule in Section 3 of Dawid and Lauritzen (2005), in which the sample space is Euclidean. Hyvärinen (2005) introduced this scoring rule (implicitly) in the context of estimation, where it serves to consistently estimate statistical models when probability density functions are known up to a multiplicative normalization constant only.

We are furthermore grateful to Philip Dawid, Steffen Lauritzen and Matt Parry for pointing out to us that the uniqueness claim in our Theorem 3.8 is false. The error in the proof lies at the end of Step 4, where we overlooked the remaining relations $a_{0l} = (l+2)b_{0,l+2}$ for $l \geq 0$. With this, subject to the regularity conditions of the theorem, a regular local proper scoring rule of order $\lambda = 2$ necessarily takes the form

$$\mathrm{S}(p,x) = \mathrm{S}(y_0, y_1, y_2) = c_0 + c_1 y_0 - \sum_{k=2}^{\infty} c_k y_1^{k-2} \left(y_1^2 + k y_2\right)$$

where $y_j = y_j(x) = (\log p)^{(j)}(x)$ for $j \geq 0$, and the c_k for $k \geq 0$ are real constants. When $c_1 > 0$ and $c_k = 0$ for $k \neq 1$, we obtain the logarithmic score. When $c_2 > 0$ and $c_k = 0$ for $k \neq 2$, the Fisher score arises. When $c_4 > 0$ and $c_k = 0$ for $k \neq 4$ we obtain a scoring rule that is proper relative to suitable classes $\mathcal{P}_0$. However, when $c_3 \neq 0$ and $c_k = 0$ for $k \neq 3$ the resulting scoring rule is improper relative to the Gaussian measures.

Additional references

Dawid, A. P. and Lauritzen, S. L. (2005). The geometry of decision theory. Proceedings of the Second International Symposium on Information Geometry and its Applications, University of Tokyo, pages 22–28.

Hyvärinen, A. (2005). Evaluation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, **6**, 695–709.

Local Proper Scoring Rules

Werner Ehm and Tilmann Gneiting

Institute for Frontier Areas of Psychology and Mental Health
University of Washington

Technical Report no. 551

Department of Statistics, University of Washington

February 6, 2009

Abstract

Scoring rules assess the quality of probabilistic forecasts, by assigning a numerical score based on the predictive distribution and on the event or value that materializes. A scoring rule is proper if it encourages truthful reporting. It is local of order λ if the score depends on the predictive density only through its value and its derivatives of order up to λ at the observation. Previously, only a single local proper scoring rule had been known, namely the logarithmic score, which is local of order $\lambda = 0$. Here we introduce the Fisher score, which is a local proper scoring rule of order $\lambda = 2$. It relates to the Fisher information in the same way that the logarithmic score relates to the Kullback-Leibler information. The convex cone generated by the logarithmic score and the Fisher score exhausts the class of the local proper scoring rules of order $\lambda \leq 2$, up to equivalence and regularity conditions. In a data example, we use local and non-local proper scoring rules to assess statistically postprocessed ensemble weather forecasts. Finally, we develop a multivariate version of the Fisher score.

1 Introduction

One of the major purposes of statistical analysis is to make forecasts for the future, and to provide suitable measures of the uncertainty associated with them. Consequently, forecasts ought to be probabilistic in nature, taking the form of probability distributions over future quantities and events (Dawid 1984; Gneiting 2008). Scoring rules provide summary measures for the evaluation of probabilistic forecasts, by assigning a numerical score based on the

AMS 2000 subject classifications. Primary 62C05; secondary 62M20, 86A10.

Key words and phrases. Bregman divergence; Density forecast; Euler equation; Fisher information; Forecast verification; Logarithmic score; Probabilistic forecasting; Proper scoring rule.

predictive distribution and on the event or value that materializes. We take scoring rules to be positively oriented rewards that a forecaster wishes to maximize. Specifically, if the forecaster quotes the predictive distribution P and the event x materializes, her reward is $\mathrm{S}(P, x)$. The function $\mathrm{S}(P, \cdot\,)$ takes values in the extended real line, $\overline{\mathbb{R}} = [-\infty, \infty]$, and we write $\mathrm{S}(P, Q)$ for the expected value of $\mathrm{S}(P, \cdot\,)$ under Q. Suppose, then, that the forecaster's best judgement is the predictive distribution Q. The forecaster has no incentive to predict any $P \neq Q$, and is encouraged to quote her true belief, $P = Q$, if

$$\mathrm{S}(Q, Q) \geq \mathrm{S}(P, Q)$$

with equality if and only if $P = Q$ (Savage 1971; Bröcker and Smith 2007; Gneiting and Raftery 2007). A scoring rule with this property is said to be strictly proper. If $\mathrm{S}(Q, Q) \geq \mathrm{S}(P, Q)$ for all P and Q, the scoring rule is proper.

If the predictive distribution is absolutely continuous on the real line, $\mathbb{R}$, it can be argued that $\mathrm{S}(P, x)$ ought to depend only on the behavior of the predictive density, p, in an infinitesimal neighborhood of the observation, x. Any such scoring rule is said to be local. Hitherto, only a single local proper scoring rule had been known, namely the logarithmic score (Good 1952; Bernardo 1979), which can be understood as a predictive likelihood. The goal of our paper is to introduce and study another local proper scoring rule, namely the Fisher score. It relates to the Fisher information in the same way that the logarithmic score relates to the Kullback-Leibler information.

The remainder of the paper is organized as follows. Section 2 introduces scoring rules from a measure theoretic perspective and discusses notions of locality. In Section 3 we present the Fisher score, which we define as

$$\mathrm{FS}(p, x) = \left(\frac{p'(x)}{p(x)}\right)^2 - 2\,\frac{p''(x)}{p(x)},$$

where p is a smooth predictive density on the real line, and x is the verifying observation. The Fisher score is local of order $\lambda = 2$, in the sense that it depends on the predictive density through its value, and its first and second derivative, at the observation only. Subject to regularity conditions, the convex cone generated by the logarithmic score and the Fisher score exhausts the class of the local proper scoring rules of order $\lambda \leq 2$. A data example on ensemble weather forecasts is given in Section 4. Section 5 introduces a multivariate version of the Fisher score, which substitutes the gradient for the first, and the Laplacian for the second derivative. The paper ends with a discussion in Section 6.

2 Notions of locality for proper scoring rules

2.1 Proper scoring rules

We consider predictive distributions on a general sample space, Ω. Let $\mathcal{A}$ be a σ-algebra of subsets of Ω, and let $\mathcal{P}$ be a convex class of probability measures on $(\Omega, \mathcal{A})$. A function

on Ω is $\mathcal{P}$-*quasiintegrable* if it is measurable with respect to $\mathcal{A}$ and quasiintegrable with respect to all $P \in \mathcal{P}$ (Bauer 2001, p. 64). A *probabilistic forecast* or a *predictive distribution* is any probability measure $P \in \mathcal{P}$. A *scoring rule* is any extended real-valued function $\mathrm{S} : \mathcal{P} \times \Omega \to \overline{\mathbb{R}}$ such that $\mathrm{S}(P, \cdot)$ is $\mathcal{P}$-quasiintegrable for all $P \in \mathcal{P}$. Hence, if the forecast is P and ω materializes, the forecaster's reward is $\mathrm{S}(P, \omega)$. We define

$$\mathrm{S}(P, Q) = \int \mathrm{S}(P, \omega) \, \mathrm{d}Q(\omega)$$

as the expected score under Q when the probabilistic forecast is P. This is a well defined extended real-valued quantity, because $S(P, \cdot)$ is quasi-integrable with respect to Q.

Definition 2.1. The scoring rule S is *proper* relative to the class $\mathcal{P}$ if

$$\mathrm{S}(Q, Q) \geq \mathrm{S}(P, Q) \quad \text{for all} \quad P, Q \in \mathcal{P}.$$

It is *strictly proper* relative to $\mathcal{P}$ if $\mathrm{S}(Q, Q) \geq \mathrm{S}(P, Q)$ with equality if and only if $P = Q$.

The term proper was coined by Winkler and Murphy (1968), while the general idea can be traced to Brier (1950) and Good (1952). See Dawid (2008) for a concise and lucid history of proper scoring rules, which includes major contributions by meteorologists and by the subjective school of probability.

It will be convenient to define an equivalence relation for scoring rules (Dawid 2007). Scoring rules that are equivalent rank competing forecasters identically, and for most purposes need not be distinguished.

Definition 2.2. The scoring rules S_1 and S_2 are *equivalent* if

$$\mathrm{S}_1(P, \omega) = a \, \mathrm{S}_2(P, \omega) + b(\omega)$$

for $P \in \mathcal{P}$ and $\omega \in \Omega$, where $a > 0$ is a constant and b is a $\mathcal{P}$-integrable function.

For characterizations of proper scoring rules, see Hendrickson and Buehler (1971) and Gneiting and Raftery (2007).

2.2 Local scoring rules

Generally, a scoring rule can be thought of as *local* if $\mathrm{S}(P, \omega)$ depends on the predictive distribution, P, only through its behavior in an infinitesimal neighborhood of the verifying observation, ω. Indeed, Bernardo (1979, p. 689) argued in this vein, noting that

> "when assessing the worthiness of a scientist's final conclusions, only the probability he attaches to a small interval containing the true value should be taken into account."

We now discuss locality in the measure theoretic context of predictive densities, which includes probability forecasts on discrete sample spaces as a special case. Specifically, let μ be a σ-finite measure on the measurable space $(\Omega, \mathcal{A})$, and let $\mathcal{L}$ denote the class of the probability measures on $(\Omega, \mathcal{A})$ which are absolutely continuous with respect to μ. We identify a probabilistic forecast $P \in \mathcal{L}$ with its μ-density, p, to which we refer as a *predictive density* or a *density forecast*. Thus we henceforth write $\mathrm{S}(p, \cdot\,)$. Predictive densities are generally defined only up to a set of μ-measure zero, and it is convenient to impose regularity conditions. If the sample space is metric, Bernardo (1979) suggests the use of the unique version defined by $p(\omega) = \lim_{\rho \to 0} P(B_\rho(\omega))/\mu(B_\rho(\omega))$, where $B_\rho(\omega)$ is a ball of radius ρ centered at ω. Without loss of generality, we assume that a predictive density is essentially bounded.

The classical example of a local proper scoring rule is the *logarithmic score*, which is defined as

$$\mathrm{LS}(p, \omega) = \log(p(\omega))$$

and can be interpreted as a predictive likelihood. It was originally proposed by Good (1952) and has been widely used since, despite harsh criticism for a lack of robustness and potentially infinite penalties (Selten 1998). The logarithmic score is strictly proper relative to the class $\mathcal{L}$ by the nonnegativity of the Kullback-Leibler or entropy divergence,

$$d_{\mathrm{LS}}(p, q) = \mathrm{LS}(q, q) - \mathrm{LS}(p, q) = \int \frac{\log(q(\omega))}{\log(p(\omega))}\, q(\omega)\, \mathrm{d}\mu(\omega) \geq 0,$$

with equality if and only if $p = q$ almost surely.[1] The associated expected score function is the Shannon entropy, up to sign, in that

$$\mathrm{LS}(p, p) = \int \log(p(\omega))\, p(\omega)\, \mathrm{d}\mu(\omega).$$

On finite sample spaces, a scoring rule is thought of as *local* if $\mathrm{S}(p, \omega) = \mathrm{s}(p(\omega))$ for some function $\mathrm{s} : \mathbb{R} \to \overline{\mathbb{R}}$, and the logarithmic scoring rule is the only nontrivial local proper score, up to equivalence (McCarthy 1956; Aczel and Pfanzagl 1966; Shuford, Albert and Massengill 1966; Savage 1971). Hereinafter, we are concerned with Euclidean sample spaces, in which distinct notions of locality are conceivable. Initially, we consider the case in which the sample space is the real line.

2.3 Local scoring rules on the real line

We turn to the familiar case in which $\Omega = \mathbb{R}$ is the real line, $\mathcal{A}$ is the Borel σ-algebra, μ is the Lebesgue measure, and $\mathcal{P}$ is some class of Borel probability measures that admit a smooth Lebesgue density, p. In this setting, we define locality as follows.

[1] Our notation follows Gneiting and Raftery (2007), who identify the first argument, p, with the forecast, and the second argument, q, with the true, data-generating distribution. Traditionally, the Kullback-Leibler divergence has been defined with the roles of p and q interchanged.

Definition 2.3. Let λ be a nonnegative integer, and let S be a scoring rule for some class $\mathcal{P}$ of Borel probability measures on $\mathbb{R}$ which admit a Lebesgue density whose derivatives up to order λ exist and are continuous. Then S is *local of order* λ if there exists a function $\mathrm{s} : \mathbb{R}^{2+\lambda} \to \overline{\mathbb{R}}$, to be called *scoring function,* such that

$$\mathrm{S}(p, x) = \mathrm{s}(x, p(x), p'(x), \ldots, p^{(\lambda)}(x)).$$

Bernardo (1979) showed that the logarithmic scoring rule is the only local proper scoring rule of order $\lambda = 0$ on the real line, up to equivalence and regularity conditions.

3 The Fisher score

Perhaps surprisingly, there are nontrivial local proper scoring rule on the real line other than the logarithmic score. Specifically, we introduce the Fisher score, which is a local proper scoring rule of order $\lambda = 2$.

3.1 Valid classes of probability measures

Before defining the Fisher score and discussing its properties, we consider relevant classes of predictive distributions with smooth Lebesgue densities.

Definition 3.1. Suppose that $\mathcal{P}$ is a class of Borel probability measures on $\mathbb{R}$ that are absolutely continuous with respect to the Lebesgue measure. We identify an element of the class $\mathcal{P}$ with its Lebesgue density, p. Let $\Omega_+ \subseteq \mathbb{R}$ be an open set. Then the class $\mathcal{P}$ is *valid* with *fundamental domain* Ω_+ if the following conditions are satisfied:

(a) For all $p \in \mathcal{P}$, $p(x) > 0$ if and only if $x \in \Omega_+$.

(b) All $p \in \mathcal{P}$ admit a continuous second derivative, p'', on $\mathbb{R}$.

(c) For all $p, q \in \mathcal{P}$,

$$\int \left(\frac{p'(x)}{p(x)}\right)^2 q(x)\,dx < \infty \qquad \text{and} \qquad \int \frac{|p''(x)|}{p(x)}\, q(x)\,dx < \infty\,. \tag{1}$$

Valid classes of probability measures enjoy a number of desirable and useful properties, including the following.

Proposition 3.2. *Any valid class of Borel probability measures is convex.*

Proof. Clearly, the properties in (a) and (b) are preserved under convex combinations, so we only need to be concerned about part (c). Let p_0, p_1 and q_0, q_1 be elements of the valid class $\mathcal{P}$, and let $\alpha \in [0,1]$. We proceed to show that if $p = (1-\alpha)p_0 + \alpha p_1$ and $q = (1-\alpha)q_0 + \alpha q_1$ then both integrals in (1) are finite. Towards this end, consider the function

$$g(\alpha) = \frac{(1-\alpha)c_0 + \alpha c_1}{(1-\alpha)d_0 + \alpha d_1}$$

where $c_0, c_1 \in \mathbb{R}$ and $d_0, d_1 > 0$ are constants. Then

$$|g(\alpha)| \leq \max(|g(0)|, |g(1)|) \leq |g(0)| + |g(1)| \quad \text{for all} \quad \alpha \in [0,1], \tag{2}$$

because

$$g'(\alpha) = \frac{c_1 d_0 - c_0 d_1}{((1-\alpha)d_0 + \alpha d_1)^2}$$

does not change sign. Applying (2) with $c_0 = p_0'(x)$, $c_1 = p_1'(x)$, $d_0 = p_0(x)$ and $d_1 = p_1(x)$, so that $g(\alpha) = p'(x)/p(x)$, we see that

$$\int \left(\frac{p'(x)}{p(x)}\right)^2 q(x)\,dx \leq 2\int \left[\left(\frac{p_0'(x)}{p_0(x)}\right)^2 + \left(\frac{p_1'(x)}{p_1(x)}\right)^2\right](q_0(x) + q_1(x))\,dx < \infty.$$

An analogous argument applies to the second integral in (1), thereby showing that property (c) is preserved under convex combinations. ∎

The valid class in the following example is very broad. For example, it contains all normal densities, all Student t densities with $\nu > 2\alpha$ degrees of freedom, all logistic densities, the respective skew densities, and finite mixtures of the above, among many others densities of applied and theoretical interest.

Example 3.3. For $\alpha > 0$, let $\mathcal{P}_\alpha$ denote the class of the absolutely continuous Borel probability measures on $\mathbb{R}$, the Lebesgue density, p, of which satisfies the following conditions:

(a) p is strictly positive on $\mathbb{R}$;

(b) p'' exists and is continuous on $\mathbb{R}$;

(c) there exists an $\epsilon > 0$ such that

$$p(x)|x|^{1+2\alpha+\epsilon}, \qquad \frac{p'(x)}{|x|^\alpha\, p(x)} \qquad \text{and} \qquad \frac{p''(x)}{|x|^{2\alpha}\, p(x)}$$

are bounded as $x \to \pm\infty$.

Then $\mathcal{P}_\alpha$ is valid with fundamental domain $\Omega_+ = \mathbb{R}$.

3.2 The Fisher score on the real line

We now define the Fisher score, which is a local proper scoring rule of order $\lambda = 2$ on the real line, $\mathbb{R}$.

Definition 3.4. Given $x \in \mathbb{R}$ and a Lebesgue density, p, that is twice continuously differentiable on $\mathbb{R}$, the *Fisher score* is defined as

$$\mathrm{FS}(p, x) = \left(\frac{p'(x)}{p(x)}\right)^2 - 2\,\frac{p''(x)}{p(x)} \tag{3}$$

if $p(x) > 0$, and $\mathrm{FS}(p, x) = -\infty$ if $p(x) = 0$.

For later use, we note that (3) can also be written as

$$\mathrm{FS}(p, x) = -((\log p)'(x))^2 - 2\,(\log p)''(x).$$

With this, we are ready to state and prove the propriety of the Fisher score.

Theorem 3.5. *Suppose that $\mathcal{P}$ is a valid class of Borel probability measures on $\mathbb{R}$ with fundamental domain Ω_+. Then the Fisher score is proper relative to $\mathcal{P}$. It is strictly proper relative to $\mathcal{P}$ if Ω_+ is an interval.*

Proof. For brevity we suppress arguments and differentials in the integrals below. We first show that if $p, q \in \mathcal{P}$ then

$$\int \left(\frac{p''}{p}\right) q \;=\; \int \left(\frac{p'}{p}\right)^2 q \;-\; \int \left(\frac{p'q'}{pq}\right) q. \tag{4}$$

Indeed, partial integration applied on the finite interval $[a, b]$ gives

$$\begin{aligned} \int_a^b p'' \left(\frac{q}{p}\right) \;&=\; p'\,\frac{q}{p}\bigg|_a^b \;-\; \int_a^b p' \left(\frac{q'}{p} - \frac{qp'}{p^2}\right) && (5)\\ &=\; \frac{p'}{p}\,q\bigg|_a^b \;+\; \int_a^b \left(\frac{p'}{p}\right)^2 q \;-\; \int_a^b \frac{p'q'}{p}. && (6) \end{aligned}$$

As $a \to -\infty$ and $b \to \infty$, the left-hand side of (5) and the second term in (6) converge to finite limits. The third term in (6) can be estimated as

$$\int_a^b \left|\frac{p'}{p}\sqrt{q}\right| \left|\frac{q'}{q}\sqrt{q}\right| \;\le\; \left[\int_{-\infty}^{\infty} \left(\frac{p'}{p}\right)^2 q\right]^{1/2} \left[\int_{-\infty}^{\infty} \left(\frac{q'}{q}\right)^2 q\right]^{1/2} \;<\; \infty.$$

Hence, the third term in (6) converges to a finite limit as well. This implies that the boundary term allows for a finite limit, in that

$$\lim_{a \to -\infty,\, b \to \infty} \left[\frac{p'(x)}{p(x)}\,q(x)\right]_{x=a}^{b} \;=\; C$$

for some $C \in \mathbb{R}$. However,

$$\int \frac{|p'|}{p}\, q \;\leq\; \left[\int \left(\frac{p'}{p}\right)^2 q \right]^{1/2}$$

is finite, so the limit value is $C = 0$ and we have demonstrated (4). It is thus immediate that

$$\begin{aligned} \mathrm{FS}(q,q) - \mathrm{FS}(p,q) &= \int \left(\frac{q'}{q}\right)^2 q - \int \left(\frac{p'}{p}\right)^2 q + 2 \int \left(\frac{p''}{p}\right) q \\ &= \int \left(\frac{q'}{q}\right)^2 q - 2 \int \left(\frac{p'q'}{pq}\right) q + \int \left(\frac{p'}{p}\right)^2 q \\ &= \int \left(\frac{q'}{q} - \frac{p'}{p}\right)^2 q. \end{aligned} \tag{7}$$

The latter integral is nonnegative, which establishes the statement about propriety. To prove the claim about strict propriety, suppose that the integral vanishes. Then for each connected component of Ω_+ there exists a constant $c \in \mathbb{R}$ such that $p = cq$ on this component. If Ω_+ is an interval, then $p = q$ on Ω_+ and hence on $\mathbb{R}$. ■

The associated expected score function or information measure (Gneiting and Raftery 2007) is the Fisher information,

$$\mathrm{FS}(p,p) = \mathrm{FI}(p) = \int \left(\frac{p'(x)}{p(x)}\right)^2 p(x)\, dx. \tag{8}$$

The Fisher information is finite if p belongs to a valid class. The respective functional Bregman divergence (Gneiting and Raftery 2007; Frigyik, Srivastava and Gupta 2008) is the Fisher information divergence (7), namely[2]

$$d_{\mathrm{FS}}(p,q) = \mathrm{FS}(q,q) - \mathrm{FS}(p,q) = \int \left(\frac{q'(x)}{q(x)} - \frac{p'(x)}{p(x)}\right)^2 q(x)\, dx.$$

Table 1 shows the Fisher score and the Fisher information for standardized normal, t, and logistic densities. The respective location-scale families satisfy

$$\mathrm{FS}\left(\frac{1}{\sigma}\, p\left(\frac{\cdot - \mu}{\sigma}\right),\, x\right) = \frac{1}{\sigma^2}\, \mathrm{FS}\left(p(\cdot), \frac{x - \mu}{\sigma}\right) \tag{9}$$

and

$$\mathrm{FI}\left(\frac{1}{\sigma}\, p\left(\frac{\cdot - \mu}{\sigma}\right)\right) = \frac{1}{\sigma^2}\, \mathrm{FI}(\, p(\cdot)). \tag{10}$$

[2]Again, our notation follows the proposal of Gneiting and Raftery (2007), which is geared to the current context. Traditionally, the Fisher information divergence has been defined with the roles of p and q interchanged.

Table 1: Fisher score and Fisher information for standardized normal, t_ν ($\nu > 0$), and logistic densities. Score and information for the associated location-scale families derive easily from (9) and (10).

Density	Fisher score	Fisher information
Normal	$2 - x^2$	1
t_ν	$(\nu+1)(2\nu - (\nu+3)x^2)/(\nu + x^2)^2$	$(\nu+1)/(\nu+3)$
Logistic	$(6e^x - e^{2x} - 1)/(1+e^x)^2$	$1/3$

The statement about strict propriety in Theorem 3.5 hinges on the additional assumption that the fundamental domain is an interval. If the fundamental domain is disconnected, there are counterexamples. For instance, consider the fundamental domain $\Omega_+ = (-1, 0) \cup (0, 1)$ and the density

$$p_\alpha(x) \;\propto\; (1-\alpha)\, x^2(1+x^2)\, \mathbb{I}_{(-1,0)}(x) + \alpha\, x^2(1-x^2)\, \mathbb{I}_{(0,1)}(x),$$

where $0 < \alpha < 1$ (Huber 1974; Huber 1981, p. 81). Then $\mathrm{FS}(p_\alpha, p_\beta) = 0$ for all $0 < \alpha, \beta < 1$, in violation of strict propriety.

3.3 Uniqueness

Along with the logarithmic and the Fisher score, any convex combination thereof is a local proper scoring rule. Does the convex cone generated by the two scores exhaust the class of the local proper scoring rules of order $\lambda \leq 2$, up to equivalence and regularity conditions, just as the ray generated by the logarithmic score does in the case $\lambda = 0$? Our goal here is to develop and state a version of such a result answering the question in the positive.

We start from the observation that in terms of the quantities

$$z_j \equiv z_j(p, x) = (\log p)^{(j)}(x) \qquad \text{where} \qquad j = 0, 1, 2, \ldots,$$

the logarithmic score and the Fisher score assume simple analytic forms, namely z_0 and $-z_1^2 - 2z_2$, respectively. This prompts us to restrict attention to the following class of local scoring rules.

Definition 3.6. For $\lambda = 0, 1, \ldots$, let $\mathcal{R}_\lambda$ denote the class of the entire (real-)analytic functions (having a globally convergent power series expansion) in $\lambda + 1$ real variables $y_0, \ldots, y_\lambda$. A local scoring rule S of order λ is *regular* if it is of the form

$$\mathrm{S}(p, x) = \mathrm{s}(z_0(p, x), \ldots, z_\lambda(p, x)) \quad \text{for some} \quad \mathrm{s} \in \mathcal{R}_\lambda. \tag{11}$$

Here we are interested in the case $\lambda = 2$. The envisaged characterization will follow from the restrictions induced by the requirement of propriety. Specifically, if the scoring rule S is proper, the functional $\mathcal{P} \ni p \longmapsto \mathrm{S}(p, q)$ achieves its maximum at $p = q$, *for every* $q \in \mathcal{P}$. An application of the calculus of variations then shows, subject to boundary conditions, that *the scoring function* $\mathrm{s} = \mathrm{s}(z_0(q, \cdot), z_1(q, \cdot), z_2(q, \cdot))$ *is such that the differential equation*

$$\partial_0 \mathrm{s} - \frac{1}{q} \frac{d}{dx} (q \cdot \partial_1 \mathrm{s}) + \frac{1}{q} \frac{d^2}{dx^2} (q \cdot \partial_2 \mathrm{s}) = c \tag{12}$$

holds for every $q \in \mathcal{P}$. Here c is the Lagrange multiplier associated with the side condition that p integrates to unity, and the notation $\partial_k = \partial / \partial z_k$ is used for convenience. One verifies readily that the logarithmic score and the Fisher score satisfy the differential equation (12) with $c = 1$ and $c = 0$, respectively.

Equation (12) essentially is the Euler equation of the calculus of variations (Gelfand and Fomin 1963, pp. 40–42). Its slightly different form here results from the fact that in our case the integrand of the functional to be optimized is of the form $F(\log y, (\log y)', (\log y)'')$ rather than of the common form $F(y, y', y'')$. Conditions for the validity of (12) will be given in Proposition 3.11 below. The intended characterization will be stated independently of these issues, under the assumption that (12) holds for all $q \in \mathcal{P}_0$, where $\mathcal{P}_0 \subseteq \mathcal{P}$ is a sufficiently large subclass.

Definition 3.7. Let $\mathcal{P}_0$ be a valid class of probability measures with fundamental domain Ω_+. Let $k \geq 0$ be an integer.

(a) We say that *condition* (B_k) *holds at* $x \in \Omega_+$ if the set $\{(z_0(p, x), \ldots, z_k(p, x)) : \ p \in \mathcal{P}_0\}$ is an open subset of $\mathbb{R}^{k+1}$.

(b) The class $\mathcal{P}_0$ is *k-rich* if every $q \in \mathcal{P}_0$ is k times continuously differentiable, condition (B_k) holds at some $x \in \Omega_+$, and furthermore

$$\inf_{p_1, p_2 \in \mathcal{P}_0} \frac{d_{\mathrm{LS}}(p_1, p_2)}{d_{\mathrm{FS}}(p_1, p_2)} = 0 \qquad \text{and} \qquad \sup_{p_1, p_2 \in \mathcal{P}_0} \frac{d_{\mathrm{LS}}(p_1, p_2)}{d_{\mathrm{FS}}(p_1, p_2)} = \infty. \tag{13}$$

With these preliminaries, we are ready to state our uniqueness result.

Theorem 3.8. *Let* S *be a regular local scoring rule of order* $\lambda = 2$ *that is proper relative to the valid class* $\mathcal{P}_0$ *with fundamental domain* $\Omega_+ \subseteq \mathbb{R}$. *Suppose that the associated scoring function* $\mathrm{s} \in \mathcal{R}_2$ *is not constant and such that every* $q \in \mathcal{P}_0$ *satisfies the differential equation* (12). *If the class* $\mathcal{P}_0$ *is* 4*-rich, then* S *is a convex combination of the logarithmic score and the Fisher score, up to equivalence.*

Note, in particular, that allowing for order $\lambda = 1$ does not suffice to extend the class of the regular local proper scoring rules beyond the logarithmic score, which is local of order $\lambda = 0$.

Our characterization of the (regular) local proper scoring rules of order $\lambda = 2$ applies under general conditions. The assumptions on the class $\mathcal{P}_0$ are weak and readily satisfied, as we demonstrate in the following examples, in which $\Omega_+ = \mathbb{R}$.

Example 3.9. Let $\mathcal{P}_0$ consist of all strictly positive densities of the form $\phi + \psi$, where ϕ is a fixed Gaussian density, and ψ is a compactly supported, infinitely differentiable function that integrates to zero. Then $\mathcal{P}_0$ is valid and 4-rich.

A larger class $\mathcal{P}_0$ that enjoys the same properties is obtained if we allow the function ϕ to vary over the class of the finite mixtures of Gaussian densities. Then $\mathcal{P}_0$ contains all Gaussian densities and condition (13) can be verified immediately.

Example 3.10. If $\mathcal{P}_0$ contains all Gaussian densities, the ratio condition (13) is satisfied. To see this, note that if $p_1 = \mathcal{N}(\mu_1, \sigma_1^2)$ and $p_2 = \mathcal{N}(\mu_2, \sigma_2^2)$ then

$$d_{\mathrm{LS}}(p_1, p_2) = \frac{1}{2}\left(\frac{(\mu_1 - \mu_2)^2}{\sigma_1^2} + \frac{\sigma_2^2}{\sigma_1^2} - 1 - \log\frac{\sigma_2^2}{\sigma_1^2}\right)$$

and

$$d_{\mathrm{FS}}(p_1, p_2) = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^4} + \frac{(\sigma_1^2 - \sigma_2^2)^2}{\sigma_1^4 \sigma_2^2}\,.$$

Evidently, the ratio $d_{\mathrm{LS}}(p_1, p_2)/d_{\mathrm{FS}}(p_1, p_2)$ can attain any positive real number as the mean and variance parameters vary.

The variational argument that yields the Euler equation (12) applies under polynomial growth and decay conditions.

Proposition 3.11. *Let $\mathcal{P}_0$ be a valid class of four times continuously differentiable probability densities. Suppose that for every $q \in \mathcal{P}_0$ and $j = 0, 1, \ldots, 4$*

$$\lim_{x \to \pm\infty} |x|^m |q^{(j)}(x)| = 0 \tag{14}$$

for all $m > 0$, and that

$$\lim_{x \to \pm\infty} \frac{|(\log q)^{(j)}(x)|}{1 + |x|^n} = 0 \tag{15}$$

for some constant n which may depend on q. Suppose furthermore that there exist finite constants C and r such that whenever v stands for $\mathrm{s} \in \mathcal{R}_2$ or any of its partial derivatives up to order 3, then

$$|v(y_0, y_1, y_2)| \le C\left[(1 + |y_0|)(1 + |y_1|)(1 + |y_2|)\right]^r \quad \text{for all} \quad (y_0, y_1, y_2) \in \mathbb{R}^3. \tag{16}$$

If the class $\mathcal{P}_0$ is complete and the scoring rule S associated with the function s is proper relative to $\mathcal{P}_0$, then the Euler equation (12) holds for every $q \in \mathcal{P}_0$.

The conditions of Proposition 3.11 do not restrict the scoring function s to be of polynomial form. For example, it could be of Hermite type, being the product of a polynomial and a Gaussian function. Example 3.9 gives an instance of a class $\mathcal{P}_0$ that satisfies the conditions of both Theorem 3.8 and Proposition 3.11.

The respective proofs are tedious and deferred to the Appendix. Briefly, our uniqueness result depends on a reduction principle for differential equations that allows for the successive simplification of the (4-th order, non-linear) Euler equation (12). Eventually, power series expansions and a study of their coefficients yield the asserted representation.

4 Data example: Probabilistic weather forecasting

The data example in this section illustrates the use of the Fisher score, along with the logarithmic score and non-local scoring rules, in an applied forecasting problem.

Weather forecasting has traditionally been viewed as a deterministic enterprise that draws on highly sophisticated, numerical models of the atmosphere. The advent of ensemble prediction systems in the early 1990s marks a change of paradigms, which has resulted in a shift towards probabilistic forecasting (Palmer 2002; Gneiting and Raftery 2005). An ensemble prediction system consists of multiple runs of numerical weather prediction models, which differ in the initial conditions being used and/or the mathematical representation of the atmosphere. Ensemble forecasts can rarely be interpreted as a random sample from the predictive distribution of future atmospheric states, in that they are subject to dispersion errors and biases. Thus some form of statistical postprocessing is required, for a happy marriage of mechanistic and statistical modeling.

Wilks (2006), Wilks and Hamill (2007) and Bröcker and Smith (2008) review various approaches to the statistical postprocessing of ensemble forecasts. Arguably, the two most prevalent methods are the Bayesian model averaging (BMA) approach developed by Raftery et al. (2005) and Sloughter et al. (2007), and the heterogeneous regression, or ensemble model output statistics (EMOS), technique of Gneiting et al. (2005). The BMA approach employs a mixture distribution, where each mixture component is a parametric probability density associated with an individual ensemble member. The mixture weights reflect the members' relative contributions to predictive skill over a training period. In contrast, the EMOS predictive distribution is a single parametric distribution.

For concreteness, consider an ensemble of point forecasts, $f_1, \ldots, f_k$, for surface temperature, x, at a given time and location. The goal is to fit predictive distributions that are as sharp as possible, subject to them being calibrated (Gneiting, Balabdaoui and Raftery 2007; Pal 2009). The BMA approach of Raftery et al. (2005) employs Gaussian components with a linearly bias-corrected mean. The BMA predictive density for temperature then becomes

$$p(x \mid f_1, \ldots, f_k) \;=\; \sum_{i=1}^{k} w_i \, \mathcal{N}(a_i + b_i f_i, \, \sigma^2),$$

Table 2: Performance of statistically postprocessed ensemble forecasts of surface temperature over the North American Pacific Northwest in April–June 2000, using Bayesian model averaging (BMA) and ensemble model output statistics (EMOS).

Scoring Rule	LS	FS	QS	SphS
BMA	−2.502	0.113	0.096	0.310
EMOS	−2.488	0.118	0.103	0.321

with BMA weights, $w_1, \ldots, w_k$, that are nonnegative and sum to 1, bias parameters $a_1, \ldots, a_k$ and $b_1, \ldots, b_k$, and a common variance parameter, σ^2. The EMOS approach of Gneiting et al. (2005) employs a single Gaussian predictive density, in that

$$p(x \mid f_1, \ldots, f_k) \;=\; \mathcal{N}(a + b_1 f_1 + \cdots + b_k f_k,\, c + d s^2),$$

with location parameters a and $b_1, \ldots, b_k$, and spread parameters c and d, where s^2 is the variance of the ensemble values. The EMOS technique thus is more parsimonious, while the BMA method is more flexible.

Following the original development in Raftery et al. (2005) and Gneiting et al. (2005), we apply the BMA and EMOS methods to the five-member University of Washington Mesoscale Ensemble over the North American Pacific Northwest (Grimit and Mass 2002), at a prediction horizon of 48 hours. Here we compare the predictive performance of the BMA and EMOS density forecasts for surface temperature verifying in the period of 24 April to 30 June 2000, which is the largest period common to those used by Raftery et al. (2005) and Gneiting et al. (2005). The predictive models were fitted on trailing training periods of length 25 days for BMA and length 40 days for EMOS, as recommended and described in the original sources. Overall, there were 23,691 individual forecast cases at individual meteorological stations and valid times, when aggregated temporally and spatially over the test period and the Pacific Northwest, comprising the US states of Washington, Oregon and Idaho, and the southern part of the Canadian province of British Columbia. All scores reported are averaged over the 23,691 forecast cases.

In Table 2, we assess these forecasts, by computing the average score under a number of proper scoring rules, including the logarithmic score and the Fisher score. In addition, we consider two well-known non-local scores, namely the quadratic score and the spherical score, which are defined as

$$\mathrm{QS}(p, x) = 2\, p(x) - \|p\|_2^2 \qquad \text{and} \qquad \mathrm{SphS}(p, x) = \frac{p(x)}{\|p\|_2},$$

where $\|\cdot\|_2$ denotes the L_2 norm. These scores are strictly proper relative to the class of the probability measures with square-integrable Lebesgue density (Matheson and Winkler

1976; Gneiting and Raftery 2007). Like all proper scores, they address both calibration and sharpness, to provide a summary measure of predictive performance.

Under all four scoring rules, the EMOS technique shows a slightly higher (that is, better) mean score than the BMA method, potentially because of its parsimony, which allows for better estimation. However, the differences pale when compared to those between the unprocessed ensemble forecast and the statistically postprocessed density forecasts. The unprocessed five-member ensemble gives a discrete predictive distribution, namely the empirical measure in $f_1, \ldots, f_5$, and thus the aforementioned scores cannot be applied directly. However, we can compute the mean scores for a smoothed ensemble forecast, which we take to be normal, with the first two moments identical to those of the empirical measure. Under this natural approach, the mean scores for the smoothed ensemble forecast are very low, reaching -21.4 for the logarithmic score, $-11,451.9$ for the Fisher score, -0.194 for the quadratic score, and 0.216 for the spherical score, which is nonnegative and thus bounded below. The strikingly low Fisher score illustrates the rule's extreme sensitivity to overconfident, underdispersed predictive distributions. This can lead to exceedingly harsh penalties, as evidenced by the scaling law (9).

5 The multivariate Fisher score

We now present a generalization of the Fisher score that applies to Euclidean sample spaces. In the framework of Section 2.1, let $\Omega = \mathbb{R}^d$, $\mathcal{A}$ the Borel σ-algebra and μ the Lebesgue measure on $\mathbb{R}^d$. For simplicity, we restrict our discussion to the case in which the fundamental domain Ω_+ consists of all of $\mathbb{R}^d$, as opposed to the broader treatment in Section 3. Accordingly, the subsequent integrals are taken with respect to the Lebesgue measure on $\mathbb{R}^d$. We write ∇ for the gradient, Δ for the Laplacian, and $|x|$ for the Euclidean norm of $x \in \mathbb{R}^d$.

The definition of a *valid* class $\mathcal{P}$ of probability measures is analogous to the one-dimensional case, with condition (1) assuming the form

$$\int \left| \frac{\nabla p(x)}{p(x)} \right|^2 q(x)\, dx < \infty \quad \text{and} \quad \int \frac{|\Delta\, p(x)|}{p(x)}\, q(x)\, dx < \infty \qquad (p, q \in \mathcal{P}).$$

Just as in the one-dimensional case, any valid class of Borel probability measures on $\mathbb{R}^d$ is convex. In analogy to Example 3.3, we can define a valid class $\mathcal{P}_\alpha$ ($\alpha > 0$) of probability densities on $\mathbb{R}^d$, for which condition (c) assumes the following form: There exist an $\epsilon > 0$ such that

$$p(x)|x|^{d+2\alpha+\epsilon}, \qquad \frac{\nabla p(x)}{|x|^\alpha\, p(x)} \qquad \text{and} \qquad \frac{\Delta p(x)}{|x|^{2\alpha}\, p(x)}$$

are bounded as $|x| \to \infty$. The class $\mathcal{P}_\alpha$ is very general and includes elliptically symmetric normal, Student's t and related densities (Holzmann, Munk and Gneiting 2006), the respective skew-elliptical densities (Genton 2004), and finite mixtures of the above, among many other densities of applied or theoretical interest.

We now define the multivariate Fisher score and demonstrate its propriety.

Definition 5.1. Given $x \in \mathbb{R}^d$ and a Lebesgue density, p, with continuous second derivatives on $\mathbb{R}^d$, the *Fisher score* is defined as

$$\mathrm{FS}(p,x) = \left|\frac{\nabla p(x)}{p(x)}\right|^2 - 2\,\frac{\Delta p(x)}{p(x)} \tag{17}$$

if $p(x) > 0$, and $\mathrm{FS}(p,x) = -\infty$ if $p(x) = 0$.

Theorem 5.2. *Suppose that $\mathcal{P}$ is a valid class of Borel probability measures with fundamental domain $\mathbb{R}^d$. Then the Fisher score is strictly proper relative to $\mathcal{P}$.*

Proof. Let dots denote inner products in $\mathbb{R}^d$. We only show that

$$\int \frac{\Delta p}{p}\,q \;=\; \int \left|\frac{\nabla p}{p}\right|^2 q - \int \left(\frac{\nabla p}{p}\cdot\frac{\nabla q}{q}\right) q \tag{18}$$

for $p, q \in \mathcal{P}$, the remainder of the proof being entirely analogous to that of Theorem 3.5. To prove (18), we let $\Gamma_r = B_r(0) \subseteq \mathbb{R}^d$ and apply the Gauss formula, resulting in

$$\begin{aligned}\int_{\Gamma_r} (\Delta p)\,\frac{q}{p} \;&=\; \int_{\partial\Gamma_r} (\nabla p\cdot n)\,\frac{q}{p} \;-\; \int_{\Gamma_r} \nabla p\cdot\nabla\left(\frac{q}{p}\right) &(19)\\ &=\; \int_{\partial\Gamma_r} (\nabla p\cdot n)\,\frac{q}{p} \;+\; \int_{\Gamma_r}\left|\frac{\nabla p}{p}\right|^2 q \;-\; \int_{\Gamma_r}\frac{\nabla p\cdot\nabla q}{p}\,, &(20)\end{aligned}$$

where n denotes the outward normal on Γ_r. Similarly to the one-dimensional case, the left-hand side of (19) and the second and third term in (20) converge to finite limits as $r \to \infty$. Consequently, the boundary integral allows for a finite limit, too, in that

$$\lim_{r\to\infty} \int_{\partial\Gamma_r} (\nabla p\cdot n)\,\frac{q}{p} \;=\; C$$

for some $C \in \mathbb{R}$. However, since $|\nabla p\cdot n| \le |\nabla p|$ and for a suitable constant c_d we have

$$c_d \int_0^\infty \left(\int_{\partial\Gamma_r} \frac{|\nabla p|}{p}\,q\right) r^{d-1}\,dr \;=\; \int \frac{|\nabla p|}{p}\,q \;\le\; \left[\int\left(\frac{|\nabla p|}{p}\right)^2 q\right]^{1/2} \;<\; \infty,$$

the limit value is $C = 0$ and we have shown (18). ∎

The expected score function or generalized entropy function associated with the multivariate Fisher score is the Fisher information,

$$\mathrm{FI}(p) = \int \left|\frac{\nabla p}{p}\right|^2 p,$$

and the respective Bregman divergence is the Fisher information divergence (DasGupta 2008, p. 26),

$$d_{\mathrm{FS}}(p,q) = \int \left| \frac{\nabla q}{q} - \frac{\nabla p}{p} \right|^2 q.$$

For other examples of proper scoring rules on $\mathbb{R}^d$, we refer to Matheson and Winkler (1976), Dawid and Sebastiani (1998), Gneiting and Raftery (2007) and Gneiting et al. (2008).

We close this section by hinting at a curious resemblance between the Fisher score and Stein's (1981) unbiased risk estimate. Specifically, let $T(x) = x + g(x)$ be an estimator of the location parameter in a standard Gaussian shift family, based on an observation $x \in \mathbb{R}^d$. Then $R(x) = d + 2\nabla \cdot g(x) + |g(x)|^2$ is an unbiased estimator of the quadratic risk of $T(x)$. If $g = \nabla \log f$ for some function $f > 0$ on $\mathbb{R}^d$, the risk estimator becomes

$$R(x) = d - \left| \frac{\nabla f(x)}{f(x)} \right|^2 + 2\,\frac{\Delta f(x)}{f(x)},$$

which may be compared to (17). It would be of interest to find any fundamental motives underlying this formal resemblance, perhaps akin to the connections drawn recently by Brown et al. (2006).

6 Discussion

With the resurgence of interest in probabilistic forecasting (Gneiting 2008), scoring rules for density forecasts are in increasing demand. A scoring rule on the real line is local of order λ if the score depends on the predictive density only through its value and its derivatives of order up to λ at the observation. We have introduced the Fisher score, which is a local proper scoring rule of order $\lambda = 2$. Arguably, this is a surprising result, in that hitherto only a single local proper scoring rule had been known, namely the logarithmic score, which is local of order $\lambda = 0$. The Fisher score generalizes to density forecasts for a vector-valued quantity, where it depends on the predictive density only through its value, gradient and Laplacian at the observation.

A different argument posits that a scoring rule for probabilistic forecasts ought to be sensitive to distance, in the sense that it rewards not just the assignment of greater mass to exactly the event or value that is observed, but also to nearby events (Staël von Holstein 1969; Jose, Nau and Winkler 2009). While either approach has appeal, locality and sensitivity to distance appear to be mutually exclusive properties, and it is not clear which one is more compelling (Mason 2008; Winkler and Jose 2008). That said, our meteorological data example seems to suggest that local and non-local proper scoring rules generally yield comparable inferences, except that local scoring rules are non-robust.

From a theoretical perspective, local proper scoring rules propose a number of intriguing problems. If the sample space is the real line, are there any regular local proper scoring

rules, other than convex combinations of the logarithmic and the Fisher score, and equivalent scores? Does this latter class comprise all local proper scoring rules of order $\lambda \leq 2$ if the condition of regularity is dropped, or if the score is allowed to depend explicitly on the verifying observation and thus is of the form $\mathrm{S}(p,x) = \mathrm{s}(x, \log p(x), (\log p)'(x), (\log p)''(x))$? We have evidence that any nonconstant local proper scoring rule of the form $\mathrm{S}(p,x) = \mathrm{s}(x, \log p(x), (\log p)'(x))$ with s real-analytic is equivalent to the logarithmic scoring rule. Therefore, the dependence on x is detached from the dependence on p, and enters via an additive term at most. Intuitively, the decoupling ought to apply generally; however, we have not been able to verify this, except for the case just mentioned. Similar questions of existence and uniqueness for local proper scoring rules arise in the multivariate case.

Appendix

Proof of Theorem 3.8

In the interest of clarity, we start with local scoring rules of general order and specialize to order $\lambda = 2$ later on. The key tool in the proof of Theorem 3.8 is the following reduction principle, which is a corollary to the uniqueness theorem for higher-order differential equations. To state it, we define $(k+1)$-dimensional ($k \geq 0$) row vectors

$$z(p,x,k) = (z_0(p,x), z_1(p,x), \ldots, z_k(p,x)) \qquad (p \in \mathcal{P}_0,\ x \in \Omega_+),$$

where as earlier we write $z_j(p,x) = (\log p)^{(j)}(x)$, and $z_j = z_j(p,\cdot)$ if the density p is fixed. Clearly then, $z_j' = z_{j+1} = z^{(j+1)}$ for $j = 0, 1, \ldots$, where the differentiation is with respect to the argument $x \in \Omega_+$.

Reduction Principle. *Let a and b be continuously differentiable functions of k real arguments. Suppose that $z_0(p,\cdot)$ is, for every $p \in \mathcal{P}_0$, a solution to the differential equation*

$$y^{(k)} \cdot a(y_0, y_0', \ldots, y^{(k-1)}) = b(y_0, y_0', \ldots, y^{(k-1)}). \tag{21}$$

If condition (B_m) *holds at $x_0 \in \Omega_+$ for some integer $m \geq k$, then $a(z(p,x_0,k-1)) = 0$ for all $p \in \mathcal{P}_0$.*

Proof. Fix $q \in \mathcal{P}_0$. Since condition (B_m) holds at $x_0 \in \Omega_+$, the set $\{z(p,x_0,m) : p \in \mathcal{P}_0\} \subseteq \mathbb{R}^{m+1}$ contains an open ball B centered at $z(q,x_0,m)$. The set

$$V = \{z_k(p,x_0) :\ p \in \mathcal{P}_0,\ z(p,x_0,k-1) = z(q,x_0,k-1)\} \subseteq \mathbb{R}$$

is a cylinder type projection of B and thus contains a neighborhood J of $z_k(q,x_0)$. Now, suppose that $a(z(q,x_0,k-1)) \neq 0$. Then there exists a neighborhood of $z(q,x_0,k-1)$ in $\mathbb{R}^k$, on which a does not vanish and b/a is continuously differentiable. Thus there is a neighborhood of x_0, within which there exists a unique solution to the equation (21), with

derivatives up to order $k-1$ at x_0 given by the components of $z(q, x_0, k-1)$. Since by assumption any $p \in \mathcal{P}_0$ is a solution to (21), uniqueness implies that the set V collapses to the point $z_k(q, x_0)$, which by the above it does not. The contradiction is resolved only if $a(z(q, x_0, k-1)) = 0$. Since $q \in \mathcal{P}_0$ was arbitrary, the proof of the reduction principle is complete.

Following these preparations, we present the proof of Theorem 3.8 in a series of five steps. Throughout, $x_0 \in \Omega_+$ denotes a point at which condition (B_4) is satisfied, as a consequence of the assumed 4-richness of the class $\mathcal{P}_0$.

Step 1 (first reduction of the Euler equation). As earlier, we write $\partial_k = \partial/\partial z_k$ for $k \geq 0$, with a similar notation for higher and mixed partial derivatives. Let $\mathrm{s} \in \mathcal{R}_2$. Explicit calculation of the total derivatives in the Euler equation (12),

$$\partial_0 \mathrm{s} - \frac{1}{q}\,\frac{d}{dx}\,(q \cdot \partial_1 \mathrm{s}) + \frac{1}{q}\,\frac{d^2}{dx^2}\,(q \cdot \partial_2 \mathrm{s}) = c,$$

results in the equation

$$\begin{aligned} c \;=\; & \partial_0 \mathrm{s} - z_1 \partial_1 \mathrm{s} - z_1 \partial_{10}^2 \mathrm{s} - z_2 \partial_{11}^2 \mathrm{s} - z_2' \partial_{12}^2 \mathrm{s} \\ & + (z_1^2 + z_2)\, \partial_2 \mathrm{s} + 2 z_1 \left(z_1 \partial_{20}^2 \mathrm{s} + z_2 \partial_{21}^2 \mathrm{s} + z_2' \partial_{22}^2 \mathrm{s} \right) \\ & + z_2 \partial_{20}^2 \mathrm{s} + z_1 \left(z_1 \partial_{200}^3 \mathrm{s} + z_2 \partial_{201}^3 \mathrm{s} + z_2' \partial_{202}^3 \mathrm{s} \right) \\ & + z_2' \partial_{21}^2 \mathrm{s} + z_2 \left(z_1 \partial_{210}^3 \mathrm{s} + z_2 \partial_{211}^3 \mathrm{s} + z_2' \partial_{212}^3 \mathrm{s} \right) \\ & + z_2'' \partial_{22}^2 \mathrm{s} + z_2' \left(z_1 \partial_{220}^3 \mathrm{s} + z_2 \partial_{221}^3 \mathrm{s} + z_2' \partial_{222}^3 \mathrm{s} \right). \end{aligned} \tag{22}$$

Closer inspection shows that equation (22) can be written in the form

$$z_0^{(4)} \cdot a(z_0, z_0', z_0'', z_0''') = b(z_0, z_0', z_0'', z_0''') \quad \text{with} \quad a = \partial_{22}^2 \mathrm{s}$$

and a suitable function b. Since condition (B_4) holds at $x_0 \in \Omega_+$, we may apply the reduction principle with $k = m = 4$ and conclude that $\partial_{22}^2 \mathrm{s}(z(q, x_0, 3)) = 0$ for every $q \in \mathcal{P}_0$. Once again by condition (B_4), the set $\{z(p, x_0, 3) : p \in \mathcal{P}_0\}$ contains a nonempty open ball $B \subseteq \mathbb{R}^4$. Since $\partial_{22}^2 \mathrm{s}$ depends on z_0, z_0' and z_0'' only, we conclude that $\partial_{22}^2 \mathrm{s}$ *vanishes on a nonempty open subset of* $\mathbb{R}^3$, namely the respective projection of the ball B.

Step 2 (affine linearity in z_2). The facts that $s \in \mathcal{R}_2$ is regular and $\partial_{22}^2 \mathrm{s}$ vanishes on a nonempty open subset of $\mathbb{R}^3$, imply that $\partial_{22}^2 \mathrm{s}$ vanishes throughout $\mathbb{R}^3$. Thus, the scoring function s admits an affine-linear representation of the form

$$\mathrm{s}(y_0, y_1, y_2) = h(y_0, y_1) + y_2\, g(y_0, y_1) \quad \text{for} \quad (y_0, y_1, y_2) \in \mathbb{R}^3, \tag{23}$$

with real-analytic coefficient functions $g, h \in \mathcal{R}_1$.

Step 3 (second reduction). The term $z_2'\partial_{12}^2 \mathrm{s} = z_2'\partial_{21}^2 \mathrm{s}$ appears both with a plus and a minus sign in (22), hence cancels. Removing all terms involving $\partial_{22}^2 \mathrm{s}$ and partial derivatives thereof, we see that the differential equation (22) simplifies to

$$\begin{aligned} c &= \partial_0 \mathrm{s} - z_1\partial_1 \mathrm{s} - z_1\partial_{10}^2 \mathrm{s} - z_2\partial_{11}^2 \mathrm{s} + (z_1^2 + z_2)\partial_2 \mathrm{s} \\ &\quad + 2z_1(z_1\partial_{20}^2 \mathrm{s} + z_2\partial_{21}^2 \mathrm{s}) + z_2\partial_{20}^2 \mathrm{s} + z_1(z_1\partial_{200}^3 \mathrm{s} + z_2\partial_{201}^3 \mathrm{s}) + z_2(z_1\partial_{210}^3 \mathrm{s} + z_2\partial_{211}^3 \mathrm{s}). \end{aligned} \tag{24}$$

Using (23) and noting that $\partial_0 s = z_2\partial_0 g + \partial_0 h$, $\partial_1 s = z_2\partial_1 g + \partial_1 h$ and $\partial_2 \mathrm{s} = g$, we can write equation (24) as

$$c = z_2\big[z_1(\partial_1 g + \partial_{01}^2 g) + g + 2\partial_0 g - \partial_{11}^2 h\big] + z_1^2(g + 2\partial_0 g + \partial_{00}^2 g) - z_1(\partial_1 h + \partial_{01}^2 h) + \partial_0 h.$$

Since $z_2 = z_0''$ and g, h depend on z_0 and $z_1 = z_0'$ only, another application of the reduction principle (with $m = 4$ and $k = 2$) yields two equations for the functions g and h, namely

$$z_1(\partial_1 g + \partial_{01}^2 g) + g + 2\partial_0 g - \partial_{11}^2 h = 0, \tag{25}$$

$$z_1^2(g + 2\partial_0 g + \partial_{00}^2 g) - z_1(\partial_1 h + \partial_{01}^2 h) + \partial_0 h = c. \tag{26}$$

Step 4 (series expansion and determination of coefficients). The functions $g, h \in \mathcal{R}_1$ admit power series representations

$$g(y_0, y_1) = \sum_{k,\ell} a_{k,\ell}\, y_0^k y_1^\ell \qquad \text{and} \qquad h(y_0, y_1) = \sum_{k,\ell} b_{k,\ell}\, y_0^k y_1^\ell.$$

By equation (25),

$$\begin{aligned} 0 &= y_1 \left(\sum_{k\ge 0,\, \ell \ge 1} \ell\, a_{k,\ell}\, y_0^k y_1^{\ell-1} + \sum_{k\ge 1,\, \ell \ge 1} k\ell\, a_{k,\ell}\, y_0^{k-1} y_1^{\ell-1} \right) + \sum_{k\ge 0,\, \ell \ge 0} a_{k,\ell}\, y_0^k y_1^\ell \\ &\quad + 2 \sum_{k\ge 1,\, \ell \ge 0} k\, a_{k,\ell}\, y_0^{k-1} y_1^\ell - \sum_{k\ge 0,\, \ell \ge 2} \ell(\ell-1)\, b_{k,\ell}\, y_0^k y_1^{\ell-2} \\ &= \sum_{k\ge 0,\, \ell \ge 1} \ell\, a_{k,\ell}\, y_0^k y_1^\ell + \sum_{k\ge 0,\, \ell \ge 1} (k+1)\ell\, a_{k+1,\ell}\, y_0^k y_1^\ell + \sum_{k\ge 0,\, \ell \ge 0} a_{k,\ell}\, y_0^k y_1^\ell \\ &\quad + 2 \sum_{k\ge 0,\, \ell \ge 0} (k+1)\, a_{k+1,\ell}\, y_0^k y_1^\ell - \sum_{k\ge 0,\, \ell \ge 0} (\ell+2)(\ell+1)\, b_{k,\ell+2}\, y_0^k y_1^\ell \\ &= \sum_{k\ge 0,\, \ell \ge 0} \Big[\ell a_{k,\ell} + (k+1)\ell a_{k+1,\ell} + a_{k,\ell} + 2(k+1)\, a_{k+1,\ell} - (\ell+2)(\ell+1)\, b_{k,\ell+2} \Big] y_0^k y_1^\ell, \end{aligned}$$

whence

$$(\ell+2)(\ell+1)\, b_{k,\ell+2} = (\ell+1)\, a_{k,\ell} + (k+1)(\ell+2)\, a_{k+1,\ell} \qquad (k \ge 0,\ \ell \ge 0). \tag{27}$$

By equation (26),

$$\begin{aligned}
c &= y_1^2\left(\sum_{k\ge0,\,\ell\ge0} a_{k,\ell}\,y_0^k y_1^\ell \;+\; 2\sum_{k\ge1,\,\ell\ge0} k\,a_{k,\ell}\,y_0^{k-1}y_1^\ell \;+\; \sum_{k\ge2,\,\ell\ge0} k(k-1)\,a_{k,\ell}\,y_0^{k-2}y_1^\ell\right)\\
&\quad - y_1\left(\sum_{k\ge0,\,\ell\ge1} \ell\,b_{k,\ell}\,y_0^k y_1^{\ell-1} \;+\; \sum_{k\ge1,\,\ell\ge1} k\ell\,b_{k,\ell}\,y_0^{k-1}y_1^{\ell-1}\right) \;+\; \sum_{k\ge1,\,\ell\ge0} k\,b_{k,\ell}\,y_0^{k-1}y_1^\ell\\
&= \sum_{k\ge0,\,\ell\ge2} a_{k,\ell-2}\,y_0^k y_1^\ell \;+\; 2\sum_{k\ge0,\,\ell\ge2} (k+1)\,a_{k+1,\ell-2}\,y_0^k y_1^\ell \;+\; \sum_{k\ge0,\,\ell\ge2} (k+2)(k+1)\,a_{k+2,\ell-2}\,y_0^k y_1^\ell\\
&\quad - \sum_{k\ge0,\,\ell\ge0} \ell\,b_{k,\ell}\,y_0^k y_1^\ell \;-\; \sum_{k\ge0,\,\ell\ge0} (k+1)\ell\,b_{k+1,\ell}\,y_0^k y_1^\ell \;+\; \sum_{k\ge0,\,\ell\ge0} (k+1)\,b_{k+1,\ell}\,y_0^k y_1^\ell\\
&= \sum_{k\ge0,\,\ell\ge2}\Big[a_{k,\ell-2} + 2(k+1)\,a_{k+1,\ell-2} + (k+2)(k+1)a_{k+2,\ell-2}\Big]y_0^k y_1^\ell\\
&\quad - \sum_{k\ge0,\,\ell\ge0}\Big[\ell\,b_{k,\ell} + (k+1)(\ell-1)\,b_{k+1,\ell}\Big]y_0^k y_1^\ell.
\end{aligned}$$

Comparing coefficients, we obtain the relationships

$$\begin{aligned}
&b_{0,0}\text{ arbitrary}, \qquad b_{1,0} = c, \qquad b_{k,0} = 0 \quad (k\ge2),\\
&b_{k,1} = 0 \quad (k\ge0)\\
&\ell\,b_{k,\ell} + (k+1)(\ell-1)\,b_{k+1,\ell}\\
&\qquad = a_{k,\ell-2} + 2\,(k+1)\,a_{k+1,\ell-2} + (k+2)(k+1)\,a_{k+2,\ell-2} \quad (k\ge0,\,\ell\ge2).
\end{aligned} \tag{28}$$

On using (27) in the form

$$\ell\,b_{k,\ell} = a_{k,\ell-2} + (k+1)\,\frac{\ell}{\ell-1}\,a_{k+1,\ell-2} \quad (k\ge0,\,\ell\ge2), \tag{29}$$

relation (28) becomes

$$(\ell-1)\,b_{k+1,\ell} = \frac{\ell-2}{\ell-1}\,a_{k+1,\ell-2} + (k+2)\,a_{k+2,\ell-2} \quad (k\ge0,\,\ell\ge2),$$

or

$$b_{k,\ell} = \frac{\ell-2}{(\ell-1)^2}\,a_{k,\ell-2} + \frac{k+1}{\ell-1}\,a_{k+1,\ell-2} \quad (k\ge1,\,\ell\ge2).$$

At the same time relation (29) holds, which may be written as

$$b_{k,\ell} = \frac{1}{\ell}\,a_{k,\ell-2} + \frac{k+1}{\ell-1}\,a_{k+1,\ell-2} \quad (k\ge0,\,\ell\ge2).$$

Comparing the preceding two relations we find that $a_{k,\ell-2}$ vanishes for $k\ge1$ and $\ell\ge2$. It then follows from (29) that $b_{k,\ell} = 0$ for $k\ge1$ and $\ell\ge2$, with $2b_{0,2} = a_{0,0}$ being the only remaining, non-trivial relation.

Step 5 (nonnegativity of the coefficients). Summarizing the development in Step 4, we find that

$$\mathrm{s}(z_0, z_1, z_2) = \alpha z_0 - \beta(z_1^2 + 2z_2) + \gamma,$$

where $\alpha = b_{1,0} = c$, $\beta = -b_{0,2} = -a_{0,0}/2$, and $\gamma = b_{0,0}$ are real constants. In other words, the function $\mathrm{s}(z_0, z_1, z_2)$ is affine in the logarithmic score, z_0, and the Fisher score, $-z_1^2 - 2z_2$. To complete the proof of Theorem 3.8, it remains to be shown that the coefficients α and β are nonnegative. This follows from propriety along with the ratio condition (13). Indeed, consider the linearly combined scoring rule $\mathrm{S}(p, x) = \alpha\mathrm{LS}(p, x) + \beta\mathrm{FS}(p, x)$. The Bregman divergence preserves linearity, and thus $d_{\mathrm{S}}(p_1, p_2) = \alpha d_{\mathrm{LS}}(p_1, p_2) + \beta d_{\mathrm{FS}}(p_1, p_2)$. By propriety, $d_{\mathrm{S}}(p_1, p_2) \geq 0$ for all $p_1, p_2 \in \mathcal{P}_0$. These facts and condition (13) are readily seen to imply that α and β are nonnegative. ∎

Proof of Proposition 3.11

The proof follows the standard route in the calculus of variations, and so we present the key points only. Fix $p, q \in \mathcal{P}_0$ and $c \in \mathbb{R}$, and consider the Lagrange function

$$L(t, c) = \int \mathrm{s}(\log p_t, (\log p_t)', (\log p_t)'')\, q \;+\; c\left(1 - \int p_t\right),$$

where $p_t = (1-t)q + tp$ for $0 \leq t \leq 1$. For brevity, we omit the argument x and the differential dx; below we furthermore suppress the arguments of s and its partial derivatives. The partial derivative of L with respect to t, when evaluated at $t = 0$, is given by

$$\begin{aligned} \partial_t L\big|_{t=0} \;=\; & \int q \left[\partial_0 \mathrm{s}(\log q, (\log q)', (\log q)'')\, \frac{p-q}{q} \right. \\ & + \partial_1 \mathrm{s}(\log q, (\log q)', (\log q)'') \left(\frac{p}{q}\right)' \\ & \left. + \partial_2 \mathrm{s}(\log q, (\log q)', (\log q)'') \left(\frac{p}{q}\right)'' \right] - c \int (p - q). \end{aligned} \tag{30}$$

We now apply the partial integration formula to the second and third terms on the right-hand side. With a view towards the boundary terms, conditions (14), (15) and (16) imply that the partial derivatives $\partial_j \mathrm{s}(\log q, (\log q)', (\log q)'')$ grow at most polynomially as functions of the (suppressed) argument x. Furthermore, after multiplication of any such expression with $p(x)$, $q(x)$ or any of their derivatives, the resulting product tends to zero as $|x| \to \infty$. Therefore, partial integration in the second term leaves no nonzero boundary terms. The integral term becomes

$$-\int p \left(\frac{1}{q}\frac{d}{dx}[q\partial_1 \mathrm{s}]\right) = -\int (p-q) \left(\frac{1}{q}\frac{d}{dx}[q\partial_1 \mathrm{s}]\right),$$

where the equality stems from the fact that

$$\lim_{b\to\infty} \int_{-b}^{b} \frac{d}{dx}[q\,\partial_1 \mathrm{s}] = \lim_{b\to\infty} \big[q\,\partial_1 \mathrm{s}\big]_{-b}^{b} = 0.$$

By twofold partial integration and similar, more tedious considerations the third term on the right-hand side of (30) can be written as

$$\int (p-q) \left(\frac{1}{q} \frac{d^2}{dx^2}[q\,\partial_2 \mathrm{s}] \right).$$

Inserting these expressions into (30) and noting that propriety implies $\partial_t L|_{t=0} = 0$, we find that

$$\int (p-q) \left[\partial_0 \mathrm{s} - \frac{1}{q} \frac{d}{dx}[q\,\partial_1 \mathrm{s}] + \frac{1}{q} \frac{d^2}{dx^2}[q\,\partial_2 \mathrm{s}] - c \right] = 0.$$

This equation holds for every $p \in \mathcal{P}_0$. By the completeness of the class $\mathcal{P}_0$, the term in brackets is constant, which gives the Euler equation up to a possibly distinct constant c. ■

Acknowledgements

The authors are grateful to Peter J. Huber for suggesting the counterexample in Section 3.2, and to Chris Fraley for providing R code used in Section 4. Tilmann Gneiting thanks the National Science Foundation for support under Awards ATM-0724721 and DMS-0706745 and the Institute for Frontier Areas of Psychology and Mental Health in Freiburg, Germany for hospitality and travel support.

References

Aczel, J. and Pfanzagl, J. (1966). Remarks on the measurement of subjective probability and information. *Metrika*, **11**, 91–105.

Bauer, H. (2001). *Measure and Integration Theory*. W. de Gruijter, Berlin.

Bernardo, J. M. (1979). Expected information as expected utility. *Annals of Statistics*, **7**, 686–690.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78**, 1–3.

Bröcker, J. and Smith, L. A. (2007). Scoring probabilistic forecasts: The importance of being proper. *Weather and Forecasting*, **22**, 382–388.

Bröcker, J. and Smith, L. A. (2008). From ensemble forecasts to predictive distribution functions. *Tellus Ser. A*, **60**, 663–678.

Brown, L., DasGupta, A., Haff, L .R. and Strawderman, W. E. (2006). The heat equation and Stein's identity: Connections, applications. *Journal of Statistical Planning and Inference*, **136**, 2254–2278.

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*. Springer, New York.

Dawid, A. P. (1984). Statistical theory: The prequential approach (with discussion and rejoinder). *Journal of the Royal Statistical Society Ser. A*, **147**, 278–292.

Dawid, A. P. (2007). The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, **59**, 77–93.

Dawid, A. P (2008). Comments on: Assessing probabilistic forecasts of multivariate quantities, with applications to ensemble predictions of surface winds. *Test*, **17**, 243–244.

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

Frigyik, B. A., Srivastava, S. and Gupta, M. R. (2008). Functional Bregman divergence and Bayesian estimation of distributions. *IEEE Transactions on Information Theory*, **54**, 5130–5139.

Gelfand, I. M. and Fomin, S. V. (1963). *Calculus of Variations*. Prentice-Hall, Englewood Cliffs, New Jersey.

Genton, M. G. (2004). *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*. Chapman & Hall/CRC, Boca Raton.

Gneiting, T. (2008). Editorial: Probabilistic forecasting. *Journal of the Royal Statistical Society Ser. A*, **171**, 319–321.

Gneiting, T. and Raftery, A. E. (2005). Weather forecasting with ensemble methods. *Science*, **310**, 248–249.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102**, 359–378.

Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Ser. B*, **69**, 243–268.

Gneiting, T., Raftery, A. E., Westveld, A. H. and Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, **133**, 1098–1118.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L. and Johnson, N. A. (2008). Assessing probabilistic forecasts of multivariate quantities, with applications to ensemble predictions of surface winds (with discussion and rejoinder). *Test*, **17**, 211–264.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society Ser. B*, **14**, 107–114.

Grimit, E. P. and Mass, C. F. (2002). Initial results of a mesoscale short-range ensemble system over the Pacific Northwest. *Weather and Forecasting*, **17**, 192–205.

Hendrickson, A. D. and Buehler, R. J. (1971). Proper scores for probability forecasters. *Annals of Mathematical Statistics*, **42**, 1916–1921.

Holzmann, H., Munk, A. and Gneiting, T. (2006). Identifiability of finite mixtures of elliptical distributions. *Scandinavian Journal of Statistics*, **33**, 753–763.

Huber, P. J. (1974). Fisher information and spline interpolation. *Annals of Statistics*, **2**, 1029–1033.

Huber, P. J. (1981). *Robust Statistics*. Wiley, New York.

Jose, V. R. R., Nau, R. F. and Winkler, R. L. (2009). Sensitivity to distance and baseline distributions in forecast evaluation. *Management Science*, in press, advance publication at DOI:10.1287/mnsc.1080.0955.

Mason, S. J. (2008). Understanding forecast verification statistics. *Meteorological Applications*, **15**, 31–40.

Matheson, J. E. and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, **22**, 1087–1096.

McCarthy, J. (1956). Measures of the value of information. *Proceedings of the National Academy of Sciences of the United States of America*, **42**, 654–655.

Pal, S. (2009). On a conjectured sharpness principle for probabilistic forecasting with calibration. *Biometrika*, in press.

Palmer, T. N. (2002). The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Quarterly Journal of the Royal Meteorological Society*, **128**, 747–774.

Raftery, A. E., Gneiting, T., Balabdaoui, F. and Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, **133**, 1155–1174.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, **66**, 783–801.

Selten, R. (1998). Axiomatic characterization of the quadratic scoring rule. *Experimental Economics*, **1**, 43–62.

Shuford, E. H., Albert, A. and Massengill, H. E. (1966). Admissible probability measurement procedures. *Psychometrika*, **31**, 125–144.

Sloughter, J. M., Raftery, A. E., Gneiting, T. and Fraley, C. (2007). Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Monthly Weather Review*, **135**, 3209–3220.

Staël von Holstein, C.-A. S. (1969). A family of strictly proper scoring rules which are sensitive to distance. *Journal of Applied Meteorology*, **9**, 360–364.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, **9**, 1135–1151.

Wilks, D. S. (2006). Comparison of ensemble-MOS methods in the Lorenz '96 setting. *Meteorological Applications*, **13**, 243–256.

Wilks, D. S. and Hamill, T. M. (2007). Comparison of ensemble-MOS methods using GFS reforecasts. *Monthly Weather Review*, **135**, 2379–2390.

Winkler, R. L. and Murphy, A. H. (1968). 'Good' probability assessors. *Journal of Applied Meteorology*, **7**, 751–758.

Winkler, R. L. and Jose, V. R. R (2008). Comments on: Assessing probabilistic forecasts of multivariate quantities, with applications to ensemble predictions of surface winds. *Test*, **17**, 251–255.