

RESEARCH ARTICLE

Forecast verification: Relating deterministic and probabilistic metrics

Tsz Yan Leung¹  | Martin Leutbecher²  | Sebastian Reich³ | Theodore G. Shepherd⁴

¹Department of Mathematics and Statistics, University of Reading, Reading, UK

²European Centre for Medium-Range Weather Forecasts, Reading, UK

³Institute of Mathematics, University of Potsdam, Potsdam, Germany

⁴Department of Meteorology, University of Reading, Reading, UK

Correspondence

T.Y. Leung, University of Reading, Department of Mathematics and Statistics, Mathematics Building, Pepper Lane, Whiteknights, Reading RG6 6AX, Berkshire, UK.
Email: t.leung@reading.ac.uk

Funding information

U.K. Engineering and Physical Sciences Research Council, Grant/Award Number: EP/L016613/1; European Research Council, Grant/Award Number: 339390; Deutsche Forschungsgemeinschaft, Grant/Award Number: SFB 1114/2/235221301

Abstract

The philosophy of forecast verification is rather different between deterministic and probabilistic verification metrics: generally speaking, deterministic metrics measure differences, whereas probabilistic metrics assess reliability and sharpness of predictive distributions. This article considers the root-mean-square error (RMSE), which can be seen as a deterministic metric, and the probabilistic metric Continuous Ranked Probability Score (CRPS), and demonstrates that under certain conditions, the CRPS can be mathematically expressed in terms of the RMSE when these metrics are aggregated. One of the required conditions is the normality of distributions. The other condition is that, while the forecast ensemble need not be calibrated, any bias or over/underdispersion cannot depend on the forecast distribution itself. Under these conditions, the CRPS is a fraction of the RMSE, and this fraction depends only on the heteroscedasticity of the ensemble spread and the measures of calibration. The derived CRPS–RMSE relationship for the case of perfect ensemble reliability is tested on simulations of idealised two-dimensional barotropic turbulence. Results suggest that the relationship holds approximately despite the normality condition not being met.

KEYWORDS

CRPS, ensembles, idealised turbulence, NWP, RMSE, verification

1 | INTRODUCTION

Operational numerical weather prediction (NWP) centres use a range of metrics to monitor and communicate forecast performance and make decisions about model upgrades. These metrics, which summarise information contained in forecasts and verifying analyses and convert them into scalar values, can broadly be divided

into two categories. The first category of metrics quantify *differences* between a single forecast and the verification field. Since these metrics depend on only one forecast state, they can be viewed as deterministic metrics. They are often used to communicate forecast skill to the public (Bauer *et al.*, 2015) as well as in theoretical predictability studies (e.g., Lorenz, 1969; Leith, 1974). On the other hand, probabilistic metrics measure the *sharpness* and *reliability* of

forecast distributions generated by ensembles of forecasts. The philosophy of probabilistic verification is therefore rather different from that of deterministic verification. That being said, it is intuitive to expect that, as probabilistic forecasts evolve in time, the loss of information manifest by the widening of forecast distributions should somehow be matched to the growth of deterministic errors when individual ensemble members, or indeed the ensemble mean, are compared against the verifying analysis. Yet not much is known about whether this relationship can be quantified mathematically, beyond the fact that the ensemble spread should agree with the root-mean-square error (RMSE) of the ensemble mean when the forecast is reliable. There has been some progress in this direction, with Gneiting and Raftery (2007) and Leutbecher and Haiden (2021) establishing certain analytic formulae for the probabilistic metric Continuous Ranked Probability Score (CRPS). In this article we shall demonstrate further that, in a bulk sense and under certain conditions, the CRPS is a function of the RMSE of the ensemble members. Furthermore, this RMSE can be related to the squared difference between the ensemble mean and the verifying analysis, which is in itself a deterministic verification metric. In this way, the CRPS–RMSE relationship may draw a link between deterministic and probabilistic verification.

The article is structured as follows. Section 2 introduces the CRPS and the RMSE, and discusses in what ways the RMSE can be interpreted as a deterministic verification metric. Our main result, the CRPS–RMSE relationship, is derived in Section 3. Its usefulness is explored for simulations of idealised two-dimensional (2D) barotropic turbulence in Section 4, where departures from the predicted relationship will be discussed in the light of the validity of the conditions imposed in the derivation. Section 5 summarises the results and concludes the article.

2 | THE METRICS

2.1 | Preliminaries

We adopt the notation of Gneiting and Raftery (2007) in respect of scoring rules for probabilistic predictions. Let P denote the predictive distribution of a scalar random variable U which materialises at value u . A scoring rule $S(P, u)$ is a function of the predictive distribution and the verification value. If, given a predictive distribution P , the verification value follows some (conditional) distribution Q , then the average score over many predictions with distribution P can be denoted by $S(P, Q) := \mathbb{E}_Q [S(P, u)]$, with

the second argument of the function $S(\cdot, \cdot)$ now being a distribution instead of a scalar value¹.

However, in contrast with the set-up of Gneiting and Raftery (2007), we shall assume that scores are negatively oriented so that forecasts with lower scores are better. Hence proper scores over a given class C of distributions have the property

$$S(Q, Q) \leq S(P, Q) \quad \forall P, Q \in C. \quad (1)$$

If, for every $Q \in C$ the equality $S(Q, Q) = S(P, Q)$ holds only when P and Q are the same², then the score is known to be strictly proper (Gneiting and Raftery, 2007). The special situation where $P = Q$ is known as the ensemble being *reliable* (although we acknowledge that other definitions and characterisations exist).

2.2 | Continuous Ranked Probability Score

The CRPS is a widely used metric that evaluates the full ensemble distribution of a continuous scalar variable and penalises unsharp distributions. It is the integral of the squared difference between the cumulative distribution function (CDF) of the forecast and of the verification:

$$\text{CRPS}(P, u) = \int_{-\infty}^{\infty} [F(x) - H_u(x)]^2 dx, \quad (2)$$

where $F(x)$ is the CDF of P and $H_u(x)$ is the Heaviside function at the verification value u .

An equivalent expression for the CRPS, often known as the ‘kernel representation’, is available for distributions P whose first moments are finite:

$$\text{CRPS}(P, u) = \mathbb{E}_P[|U - u|] - \frac{1}{2} \mathbb{E}_P[|U - U'|], \quad (3)$$

where U and U' are independent random variables drawn from the distribution P (Gneiting and Raftery, 2007). A proof of equivalence is provided in Lemmata 2.1 and 2.2 of Baringhaus and Franz (2004).

Gneiting and Raftery (2007) noted that the CRPS is a strictly proper score over a very general class of distributions, namely the class of Borel probability measures whose first moments are finite. For the special case of normal distributions $P = \mathcal{N}(\mu_P, \sigma_P^2)$, an explicit formula for

¹Without ambiguity, $S(\cdot, \cdot)$ can mean either the score for an individual prediction or the expected score over many predictions, depending on the second argument being a scalar variable or a distribution.

²This should be understood in the sense of measure theory, that P and Q only have to be equal up to a null set.

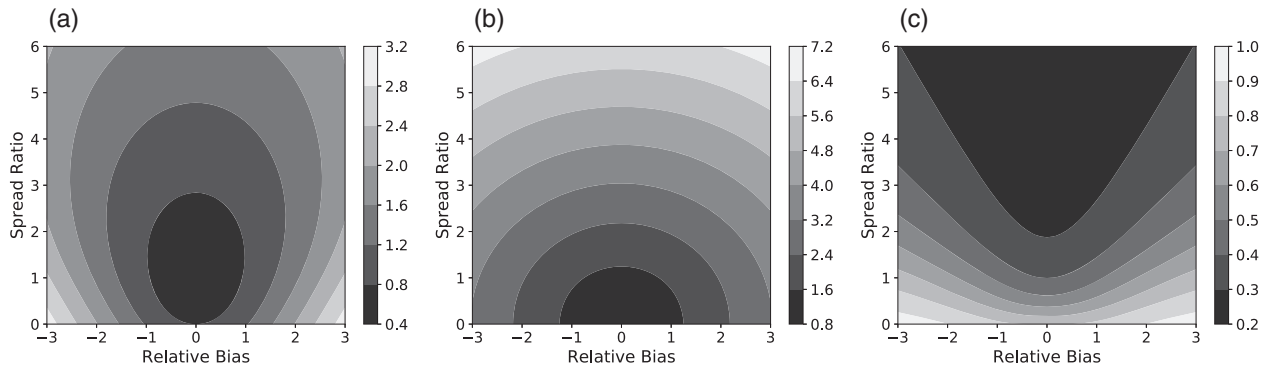


FIGURE 1 (a) $(1/\sigma_Q)\text{CRPS}(P, Q)$, (b) $(1/\sigma_Q)\text{RMSE}(P, Q)$ and (c) $\text{CRPS}(P, Q)/\text{RMSE}(P, Q)$ as functions of relative bias $b = (\mu_P - \mu_Q)/\sigma_Q$ and spread ratio $r = \sigma_P/\sigma_Q$

$\text{CRPS}(P, u)$ is available (Gneiting and Raftery, 2007):

$$\text{CRPS}(P, u) = \frac{\sigma_P}{\sqrt{\pi}} \left[-1 + \sqrt{\pi} \frac{u - \mu_P}{\sigma_P} \text{erf} \left(\frac{u - \mu_P}{\sqrt{2}\sigma_P} \right) + \sqrt{2} \exp \left\{ -\frac{1}{2} \left(\frac{u - \mu_P}{\sigma_P} \right)^2 \right\} \right], \quad (4)$$

where $\text{erf}(z) := (2/\sqrt{\pi}) \int_0^z e^{-y^2} dy$ is the error function. Denoting

$$\varphi(x) := \frac{1}{\sqrt{2\pi}} \exp \left(-\frac{1}{2}x^2 \right)$$

for the probability density function (PDF) of a standard normal random variable and

$$\Phi(x) := \int_{-\infty}^x \varphi(x') dx'$$

for its CDF, this formula could be obtained by substituting

$$F(x) = \Phi \left(\frac{x - \mu_P}{\sigma_P} \right)$$

into Equation (2), integrating by parts and invoking the identity

$$\text{erf} \left(\frac{z}{\sqrt{2}} \right) = 2\Phi(z) - 1. \quad (5)$$

Equation (4) can be integrated over a normal $\mathcal{N}(\mu_Q, \sigma_Q^2)$ kernel to yield a formula for the expected score $\text{CRPS}(P, Q)$:

$$\begin{aligned} \text{CRPS}(P, Q) &= \int_{-\infty}^{\infty} \text{CRPS}(P, u) \varphi \left(\frac{u - \mu_Q}{\sigma_Q} \right) \frac{1}{\sigma_Q} du \\ &= \int_{-\infty}^{\infty} \sigma_Q \frac{r^2}{\sqrt{\pi}} \left[-1 + \sqrt{\pi} x \text{erf} \left(\frac{x}{\sqrt{2}} \right) + \sqrt{2} \exp \left(-\frac{1}{2}x^2 \right) \right] \varphi(rx + b) dx, \quad (6) \end{aligned}$$

where

$$b := \frac{\mu_P - \mu_Q}{\sigma_Q} \quad (7)$$

is the relative bias and

$$r := \frac{\sigma_P}{\sigma_Q} \quad (8)$$

is the ratio of standard deviations, or simply the spread ratio. The Appendix demonstrates that the integral can be expressed analytically as

$$\text{CRPS}(P, Q) = \sigma_Q f(b, r), \quad (9)$$

where

$$\begin{aligned} f(b, r) &= -\frac{r}{\sqrt{\pi}} + \sqrt{\frac{2(1+r^2)}{\pi}} \exp \left(-\frac{b^2}{2(1+r^2)} \right) \\ &\quad + b \text{erf} \left(\frac{b}{\sqrt{2(1+r^2)}} \right). \quad (10) \end{aligned}$$

Note that, provided the verifying distribution Q is fixed, the qualitative properties of $\text{CRPS}(P, Q)$ are fully determined by the function $f(b, r)$ which is shown in Figure 1a. This formula for $\text{CRPS}(P, Q)$, agrees exactly with the one obtained by Leutbecher and Haiden (2021), who used the kernel representation of the CRPS (Equation (3)) as the starting point of their derivation.

2.3 | Root-mean-square error

The root-mean-square error (RMSE) is the square root of the ensemble members' mean squared error (MSE) from the verification value. The latter is defined as

$$\text{MSE}(P, u) := \mathbb{E}_P [(U - u)^2] \quad (11)$$

for an outcome $u \in \mathbb{R}$ and a distribution P of its forecast U . Mathematically speaking, this is the MSE of u as an estimator of the ensemble mean, although this could somewhat be counter-intuitive in a forecasting context. Nevertheless, the standard bias-variance decomposition of MSE applies:

$$\text{MSE}(P, u) = (\mu_P - u)^2 + \sigma_P^2, \quad (12)$$

where μ_P and σ_P are respectively the mean and the standard deviation of P . Assuming that the verifying distribution Q for u has mean μ_Q and standard deviation σ_Q , the expected score $\text{MSE}(P, Q)$ is

$$\begin{aligned} \text{MSE}(P, Q) &= \mathbb{E}_Q [(\mu_P - u)^2 + \sigma_P^2] \\ &= \sigma_Q^2 + (\mu_P - \mu_Q)^2 + \sigma_P^2 \\ &= \sigma_Q^2 (1 + b^2 + r^2), \end{aligned} \quad (13)$$

where b and r are as in Equations (7) and (8). The second equality can be established by observing that $\mathbb{E}_Q [(\mu_P - u)^2]$ is the MSE of μ_P as an estimator of u , whence the same bias-variance decomposition applies. From Equation (13), it follows that

$$\text{RMSE}(P, Q) = \sqrt{\text{MSE}(P, Q)} = \sigma_Q \sqrt{1 + b^2 + r^2}. \quad (14)$$

Note that we have not defined $\text{RMSE}(P, u)$. Should it be defined by taking the square root of Equation (11), then the $\text{RMSE}(P, Q)$ defined in Equation (14) would generally not be equal to $\mathbb{E}_Q [\text{RMSE}(P, u)]$. Hence, strictly speaking, the RMSE does not fit into the framework of scoring rules. It is simply a convenient transformation of the scoring rule $\text{MSE}(P, u)$, since it has the same physical dimensions as the variable u of interest. Given that the RMSE relates with the MSE bijectively and monotonically, we may nevertheless apply the concepts of scoring rules to the RMSE, bearing in mind that in this sense the two quantities are synonymous. The RMSE is not a proper score over any non-degenerate class of distributions (Gneiting and Raftery, 2007), as graphically confirmed in Figure 1b.

The RMSE discussed here should not be confused with the RMSE of the ensemble mean, which is based on the MSE of the ensemble mean, defined as

$$\text{MSE}_{\text{mean}}(P, u) := (\mathbb{E}_P[U] - u)^2 = (\mu_P - u)^2. \quad (15)$$

By verifying the ensemble mean as if it were a deterministic forecast in itself, $\text{MSE}_{\text{mean}}(P, u)$ and therefore its associated RMSE can be seen as a score with deterministic roots. Compared with Equation (12), $\text{MSE}_{\text{mean}}(P, u)$ lacks the contribution from the ensemble variance σ_P^2 . If the alternative formulation were to be used in place of the MSE and RMSE defined in Equations (13) and (14),

then all expressions involving $1 + b^2 + r^2$ throughout this article would have to be replaced by $1 + b^2$. (As a consequence, the multiplicative factor $\sqrt{2\pi}$ often mentioned in the forthcoming sections would become $\sqrt{\pi}$.)

3 | DERIVATION OF THE CRPS-RMSE RELATIONSHIP

So far we have seen the basic mathematical properties of the CRPS and the RMSE. Since the former is proper while the latter is improper, it is generally impossible to draw a one-to-one correspondence between the two. Nevertheless, if we compare Equations (9) and (14), we obtain

$$\text{CRPS}(P, Q) = \frac{f(b, r)}{\sqrt{1 + b^2 + r^2}} \text{RMSE}(P, Q). \quad (16)$$

What Equation (16) suggests is that, on average, the CRPS and the RMSE are related through a multiplicative factor dependent on b and r as far as predictions of normally distributed scalar variables are concerned. This multiplicative factor as a function of b and r is shown in Figure 1c. The average, as defined in Subsection 2.1, refers to aggregation over a large number of cases that share the same predictive and verifying distributions (P and Q). However, standard verification practice in NWP aggregates these scores across dimensions defined *a priori* such as grid points and forecast start dates, rather than by predictive and verifying distributions. How can Equation (16) be modified to accommodate this?

It is important to bear in mind that, in the notation $S(P, Q)$ for a given score S , there is an implied conditioning on the predictive distribution being P , since $S(P, Q)$ is the average of $S(P, u)$ over many P -distributed predictions. Q in this notation refers to the distribution of the verification value u , but it is also conditional upon the predictive distribution being P . To derive a formula for an aggregated score that takes into account the different possibilities of predictive distributions, it is therefore necessary to include information about the heteroscedasticity of P , that is, the relative frequency of occurrence of different predictive distributions. Since we have assumed that P is normal, such heteroscedasticity can be interpreted as a joint meta-distribution Θ of the parameters μ_P and σ_P . In this case the aggregated score S^* can be written as

$$S^* = \mathbb{E}_\Theta[S(P, Q)]. \quad (17)$$

This is the expectation of a conditional quantity, $S(P, Q)$. Without prescribing any specific forecast frequency distribution Θ , we can only simplify this expression further by making an extra assumption that $S(P, Q)$ be in

fact unconditional on P . This is equivalent to saying that all forecasts have the same relative bias b and spread ratio r , regardless of μ_P and σ_P . It includes the case where all forecasts are reliable (i.e., $P = Q$, or $(b, r) = (0, 1)$), but also includes the more general case where forecasts are consistently biased or over/underdispersive by a certain percentage. Substituting CRPS for S and using Equation (9), we have

$$\text{CRPS}^* = f(b, r) \mathbb{E}_\Theta [\sigma_Q] = \frac{f(b, r)}{r} \mathbb{E}_\Theta [\sigma_P]. \quad (18)$$

Similarly, substituting MSE for S and using Equation (13) gives

$$\text{MSE}^* = (1 + b^2 + r^2) \mathbb{E}_\Theta [\sigma_Q^2] = \frac{1 + b^2 + r^2}{r^2} \mathbb{E}_\Theta [\sigma_P^2] \quad (19)$$

so that

$$\text{RMSE}^* = \sqrt{\text{MSE}^*} = \frac{1}{r} \sqrt{(1 + b^2 + r^2) \mathbb{E}_\Theta [\sigma_P^2]}. \quad (20)$$

Equations (18) and (20) thus provide expressions for the CRPS and the RMSE aggregated under heteroscedastic conditions, where P 's parameters can vary from grid point to grid point, and from one forecast start date to another. These expressions assume the normality of forecast distributions as well as the consistency of the relative bias and spread ratio across all forecasts. Combining these expressions and denoting

$$h := \frac{\text{Var}_\Theta [\sigma_P]}{(\mathbb{E}_\Theta [\sigma_P])^2} \quad (21)$$

for the relative heteroscedasticity of the ensemble's standard deviation, we have

$$\begin{aligned} \frac{\text{CRPS}^*}{\text{RMSE}^*} &= \frac{f(b, r)}{\sqrt{1 + b^2 + r^2}} \frac{\mathbb{E}_\Theta [\sigma_P]}{\sqrt{\mathbb{E}_\Theta [\sigma_P^2]}} \\ &= \frac{f(b, r)}{\sqrt{1 + b^2 + r^2}} \frac{1}{\sqrt{1 + h}}. \end{aligned} \quad (22)$$

Here, we see that the ratio between the aggregated CRPS and the aggregated RMSE is the product of two terms: $f(b, r)/\sqrt{1 + b^2 + r^2}$, which depends only on the relative bias and the spread ratio; and $1/\sqrt{1 + h}$, which depends only on the relative heteroscedasticity of the ensemble spread. Since $h \geq 0$ by definition, it follows that $f(b, r)/\sqrt{1 + b^2 + r^2}$ is the upper bound of such CRPS–RMSE ratio (provided that the predictive and verifying distributions are both normal), which is the same as the multiplicative factor given in Equation (16) and shown graphically in Figure 1c. In the limit where the standard

deviation σ_P is homoscedastic, that is, $\text{Var}_\Theta [\sigma_P] \rightarrow 0$, the bound is attained and Equation (16) is recovered.

Using Equation (10), we can see that for reliable predictions of normally distributed random variables the CRPS–RMSE relationship simplifies to

$$\frac{\text{CRPS}^*}{\text{RMSE}^*} = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{1 + h}}, \quad (23)$$

which is bounded above by $1/\sqrt{2\pi}$. The bound is robust to small biases, since it is an even function with respect to the $b = 0$ axis. For example, a 5% bias only increases this bound by 0.06%. It is more sensitive to small degrees of non-calibration of the ensemble spread. To first order, the bound increases by 0.5% for every 1% of under-dispersiveness of the ensemble, and vice versa. Table 1 provides more detail on how the bound responds to small departures from perfect ensemble reliability.

4 | THE CRPS–RMSE RELATIONSHIP IN AN IDEALISED 2D TURBULENCE MODEL

The CRPS–RMSE relationship for normally distributed random variables is numerically tested in an experiment involving 2D barotropic turbulence. Due to the nature of perfect-model idealised turbulence simulations, we are only able to test the relationship for reliable predictions (Equation (23)), where $P = Q$. The turbulence is governed by the equation

$$\frac{\partial \theta}{\partial t} + J(\psi, \theta) = f + d, \quad \theta = \Delta \psi, \quad (24)$$

where t is the time, ψ is the velocity streamfunction³, θ is the vorticity, Δ is the 2D Laplacian operator⁴ and

$$J(A, B) = \frac{\partial A}{\partial x} \frac{\partial B}{\partial y} - \frac{\partial A}{\partial y} \frac{\partial B}{\partial x}.$$

Equation (24) is solved pseudo-spectrally in a doubly periodic domain, with a truncation wavenumber of $k_t = 1024$. This is equivalent to a $(2k_t) \times (2k_t) = 2048 \times 2048$ grid. The forcing f and dissipation d are prescribed in spectral space. By forcing at both large and small scales, a hybrid $k^{-3} - k^{-\frac{5}{3}}$ background spectrum is obtained, where k is the scalar wavenumber. The length-scale at which the spectral break sits respects the canonical hybrid spectrum observed and simulated in the midlatitude upper troposphere (Nastrom and Gage, 1985; Judt,

³The velocity streamfunction ψ is related to the velocity (u, v) by $u = -\partial \psi / \partial y$ and $v = \partial \psi / \partial x$.

⁴ $\Delta = \nabla \cdot \nabla$, where $\nabla = (\partial / \partial x, \partial / \partial y)$.

TABLE 1 Relative changes of $f(b, r)/\sqrt{1 + b^2 + r^2}$ compared to the case $(b, r) = (0, 1)$, when (a) the relative bias b is varied but the spread ratio r is fixed at 1, and (b) the spread ratio r is varied but the relative bias b is fixed at 0

(a)											
b	0.00	± 0.01	± 0.02	± 0.03	± 0.04	± 0.05					
Change (%)	0	+0.00	+0.01	+0.02	+0.04	+0.06					
(b)											
r	0.95	0.96	0.97	0.98	0.99	1.00	1.01	1.02	1.03	1.04	1.05
Change (%)	+2.60	+2.06	+1.53	+1.02	+0.50	0	−0.50	−0.99	−1.47	−1.94	−2.41

2020). Further details of the forcing and dissipation terms are described in Leung *et al.* (2020).

4.1 | Experimental design

A long control integration of Equation (24) is taken as the verification. When the turbulence is fully developed and reaches a statistically stationary state, a normally distributed random variable centred at zero is added to all Fourier coefficients of the vorticity field to generate the ‘truth’. The variance of the random variable, β^2 , depends on the magnitude but not the direction of the wavevector. It can be shown that a perturbation of magnitude $\rho(k)$ relative to the energy spectral density $E(k)$ of the control integration can be generated by choosing $\beta^2 = [\rho(k)E(k)/2\pi]k$. In this experiment, $\rho(k)$ is fixed to be 10^{-6} across all k . Next, $M = 4$ ensemble members are generated from the ‘truth’ using the same perturbation statistics as the generation of the ‘truth’ from the control integration. All perturbations for the four ensemble members and for the ‘truth’ are mutually independent. The perturbed simulations are integrated for a fixed time period of $T = 150$ non-dimensional units, allowing the error to almost fully saturate by the end of it.

The experiment is repeated for $N = 30$ start dates. This can be thought of as $N_1 = 5$ years, among which the control integrations are fully independent, and $N_2 = N/N_1 = 6$ start dates per year initialised at intervals of $0.1T$.

The choice of a relatively small M and large N is motivated by Leutbecher (2019). That article suggests that if the CRPS for reliable ensembles is adjusted using

$$\text{CRPS}_{\infty}^* := \frac{M}{M+1} \text{CRPS}^* \quad (25)$$

to remove the effects of the ensemble size being finite, then a reduction in the number of ensemble members used for numerical experimentation returns more robust results

than a reduction in the number of start dates, provided that the constraints in computational cost are similar.

The experimental design guarantees a reliable ensemble, since the verification is statistically indistinguishable from the M ensemble members. As such, Equation (23) is expected to hold subject to P being a normal distribution as the simulation evolves.

The scalar variables of interest chosen for this study are the velocity components u and v . For each start date and grid point, the CRPS and the MSE are computed for both velocity components in physical space. The computation of the CRPS is performed using the algorithm set out by Hersbach (2000). These metrics are then aggregated over $\Lambda := \mathcal{G} \times S \times \mathcal{D}$, where \mathcal{G} represents the set of 2048^2 grid points, S the 30 start dates and \mathcal{D} the two canonical directions (u and v), but remain as functions of the forecast lead time. Isotropy of the turbulence enables the scores for u and v to be combined without changing the results.

When the metrics are aggregated, the quantity $\mathbb{E}_{\Lambda}[S(P, u_i)]$ is computed for each lead time, where S can be CRPS or MSE, and where u_i represents a generic velocity component. The law of iterated expectations guarantees

$$\mathbb{E}_{\Lambda}[S(P, u_i)] = \mathbb{E}_{\Theta}[\mathbb{E}_P[S(P, u_i)]] = \mathbb{E}_{\Theta}[S(P, P)] = S^*, \quad (26)$$

the last two equalities of which result from the definition of S and Equation (17) respectively. In this way, CRPS^* and RMSE^* (the square-root of MSE^*) can be empirically computed, which should satisfy Equation (23) subject to the normality assumption. To account for the finite ensemble size, the aggregated CRPS is corrected by Equation (25) before being compared with the aggregated RMSE.

4.2 | Results

For notational purposes in this subsection, we denote the start date by t_0 , and write $U(t, t_0, \mathbf{x}, \mathbf{e}_1)$ for $u(t, t_0, \mathbf{x})$

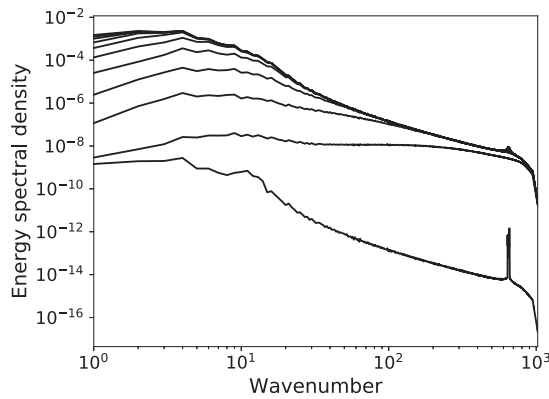


FIGURE 2 Growth of the ensemble-mean error energy spectrum, or equivalently the power spectrum of RMSE* (curves from bottom to top, plotted at intervals of $0.1T$), whose initial condition is indicated by the lowest curve

and $U(t, t_0, \mathbf{x}, \mathbf{e}_2)$ for $v(t, t_0, \mathbf{x})$. A subscript “ P ” attached to $U(t, t_0, \mathbf{x}, \mathbf{e}_i)$, $u(t, t_0, \mathbf{x})$ or $v(t, t_0, \mathbf{x})$ (where $i = 1$ or 2) indicates a forecast, in which case the variable is understood to be a random variable with distribution P . The absence of the subscript indicates the verification, which is also interpreted as a random variable but with distribution Q .

Figure 2 illustrates the growth of the error energy spectrum. More precisely, it is the spectrum of the ensemble-mean error energy aggregated over all grid points and start dates, that is, the spectral decomposition of

$$\mathbb{E}_{\mathcal{G} \times \mathcal{S}} \left[\mathbb{E}_P \left[\frac{1}{2} (\{u_f(t, t_0, \mathbf{x}) - u(t, t_0, \mathbf{x})\}^2 + \{v_f(t, t_0, \mathbf{x}) - v(t, t_0, \mathbf{x})\}^2) \right] \right]. \quad (27)$$

In two spatial dimensions⁵ and where the ensemble is reliable ($P = Q$), this is equivalent to the spectral decomposition of

$$\begin{aligned} \mathbb{E}_{\mathcal{G} \times \mathcal{S} \times \mathcal{D}} [\mathbb{E}_P [\{U_f(t, t_0, \mathbf{x}, \mathbf{e}_i) - U(t, t_0, \mathbf{x}, \mathbf{e}_i)\}^2]] \\ = \mathbb{E}_\Lambda [\text{MSE}(P, U)] = \text{MSE}^* = \text{RMSE}^{*2}, \end{aligned} \quad (28)$$

where Equations (11) and (26) have been used in the first two equalities respectively. As such, Figure 2 may also be interpreted as the evolution of the power spectrum of RMSE*. Following an initial period of adjustment that leads to fast saturation of the mesoscale $k^{-5/3}$ range, a synoptic-scale peak emerges in the error spectrum. The

spectrum then grows more or less uniformly in spatial scale and gradually saturates the k^{-3} range. After that, the growth slows down as the largest scales approach saturation. These observations are consistent with those reported in Leung *et al.* (2020).

Like RMSE*, it is possible to spectrally decompose CRPS*. To compute CRPS* for a wavenumber or range of wavenumbers, one simply picks out the associated waves in spectral space, transforms them to physical space, then aggregates the score over Λ and applies Equation (25). Such CRPS* may be compared with RMSE* for the same wavenumber(s) using Equation (23). Here, the verification metrics are decomposed into the planetary scale ($k \in [1, 8]$), synoptic scale ($k \in [9, 64]$), mesoscale ($k \in [65, 512]$) and sub-mesoscale ($k \in [513, 1024]$). The evolution of these metrics is shown in Figure 3a. Generally speaking, they grow steadily (in exponential terms) before asymptoting smoothly to their respective saturation values. The same figure also shows RMSE* associated with these scales but normalised by $\sqrt{2\pi}$ so that, according to Equation (23), the curves for the CRPS and the RMSE would coincide if P were normal and σ_P were homoscedastic. Broadly speaking, the agreement between the two is extremely close for all four spectral ranges, spanning several orders of magnitude of growth and, importantly, capturing the differences in saturation times between the different spectral ranges. This shows that, for these simulations, the normalised RMSE represents a good proxy for the CRPS. However, the discrepancy between the two curves is non-trivial throughout most of the simulation, although it remains within a factor of two. To enable closer examination of the discrepancy, the ratio of the two curves is plotted and shown in Figure 3b. Evidently, the discrepancy is stronger at the planetary and synoptic scales. For smaller scales, the CRPS and the normalised RMSE agree better, especially after the error at these scales has saturated.

Figure 4 shows the ratio $\sqrt{2\pi} \text{CRPS}^*(t)/\text{RMSE}^*(t)$ for the full field without decomposition into wavebands (thick solid curve). In addition, the thin solid curves of a lighter shade show the evolution of the ratios for the $N = 30$ individual start dates, that is, with $\Lambda = \mathcal{G} \times \mathcal{D}$ instead of $\mathcal{G} \times \mathcal{S} \times \mathcal{D}$. Considerable variation in this ratio across the 30 cases is seen, particularly at smaller lead times. According to Equation (23), the solid curves in Figure 4 are expected to coincide with

$$\frac{\mathbb{E}_\Theta [\sigma_P]}{\sqrt{\mathbb{E}_\Theta [\sigma_P^2]}} = \frac{1}{\sqrt{1+h}}$$

if the ensemble is normally distributed. Computing this ratio involves evaluating the ensemble’s standard

⁵The equivalence between Expressions (27) and (28) is not extendable to higher spatial dimensions, because it only happens in two dimensions that the factor $\frac{1}{2}$ for the kinetic energy is also the factor used to compute the average over D . In higher dimensions, the ensemble-mean error energy can be related to the MSE of velocity components via a constant multiplicative factor.

FIGURE 3 (a) $\text{CRPS}_\infty^*(t)$ (solid) and $(1/\sqrt{2\pi})\text{RMSE}^*(t)$ (dashed) for the planetary, synoptic, meso- and sub-mesoscale (from dark to light shades), up to $T = 150$. (b) The ratio $\sqrt{2\pi} \text{CRPS}_\infty^*(t)/\text{RMSE}^*(t)$ between the solid and dashed curves of (a) for the respective shades

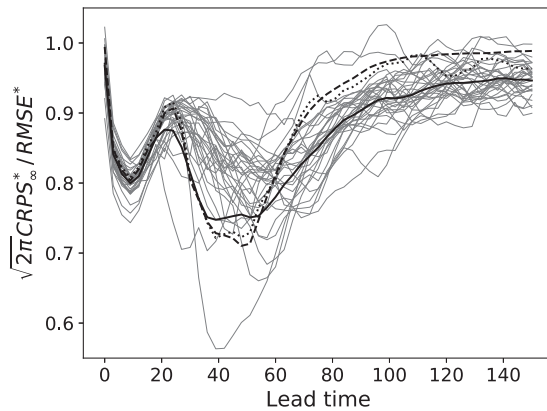
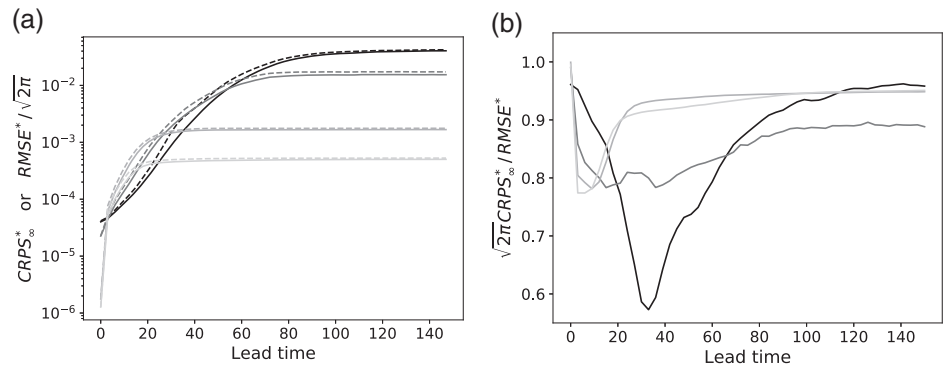


FIGURE 4 The thick solid curve is as Figure 3b, but for the full field without the scale-decomposition. The dashed curve shows $1/\sqrt{1+h}$ as a function of t for the $M = 48$ -member ensemble. The other curves show $\sqrt{2\pi} \text{CRPS}_\infty^*(t)/\text{RMSE}^*(t)$ but for $\Lambda = \mathcal{G} \times \mathcal{D}$ (i.e., for specific start dates), for the $M = 4$ -member ensemble (thin solid curves of a lighter shade) and the $M = 48$ -member ensemble (dotted curve)

deviation σ_P , but the sample size ($M = 4$) is too small to estimate σ_P robustly. To mitigate this, a larger ensemble of $M = 48$ members is run to estimate the heteroscedasticity of the ensemble's standard deviation. This is done only for a single start date ($N = 1$) owing to limited computational resources. As shown in the dashed curve of Figure 4, the fraction $1/\sqrt{1+h}$ exhibits two local minima throughout the integration, the more extreme of which corresponds to a relative heteroscedasticity of $h \approx 1$. This curve agrees nicely with the ratio $\sqrt{2\pi} \text{CRPS}_\infty^*(t)/\text{RMSE}^*(t)$ for the same large-ensemble experiment (dotted curve), which is visibly indistinguishable from the collection of thin solid curves that depict this ratio for individual cases involving the smaller ensemble ($M = 4$). The curves representing $\sqrt{2\pi} \text{CRPS}_\infty^*(t)/\text{RMSE}^*(t)$ and $1/\sqrt{1+h}$ (i.e., dotted and dashed) agree nicely, hence suggesting that heteroscedasticity is responsible for the discrepancy between $\text{CRPS}_\infty^*(t)$ and $(1/\sqrt{2\pi})\text{RMSE}^*(t)$. Any non-normality that the ensemble might develop throughout the simulation

would thus appear to have a negligible impact on the CRPS–RMSE ratio, at least for the particular case considered in this experiment.

4.3 | Non-normality of the ensemble distribution

Despite the fact that the departure of the normalised CRPS–RMSE ratio from unity can be primarily explained by the flow's heteroscedasticity, it is of interest to check explicitly whether the $M = 48$ -member ensemble is normally distributed. This is done by evaluating the ensemble's skewness and excess kurtosis at each of the 2048^2 grid points and for each of the two velocity components, and comparing histograms of these statistics across the $2048^2 \times 2$ samples with those obtained via a Monte-Carlo simulation involving $2048^2 \times 2$ groups of 48 standard normal random variables, all mutually independent. Figure 5 shows the result for several lead times. Initially the two distributions are almost identical. This can be expected, since the perturbations are normally distributed by design (Subsection 4.1). The difference grows as the flow evolves. This suggests that non-normality in the ensemble distribution is being built up as the simulation progresses, which is hardly surprising, since it is a known feature of 2D turbulence (Farge *et al.*, 1999). Yet, it also highlights that the extent of non-normality found here does not substantially affect the derived CRPS–RMSE relationship.

5 | DISCUSSION AND SUMMARY

In this article, we have derived a functional relationship between two forecast verification metrics: the CRPS and the RMSE (Sections 2 and 3). The CRPS is a standard probabilistic score that rewards forecasts that are both sharp and reliable. On the other hand, the RMSE is the sum of the ensemble variance and the squared error of the ensemble mean. In some contexts, only the latter contribution

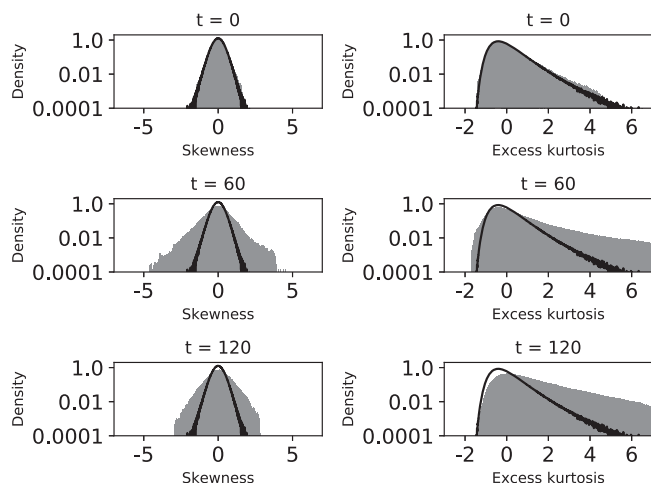


FIGURE 5 Histograms, taken over $\Lambda = G \times D$, of ensemble skewness (left) and excess kurtosis (right) of velocity components in the $M = 48$ -member ensemble simulation of idealised 2D turbulence. These are shown for lead times $t = 0$ (top), $t = 60 = 0.4T$ (middle) and $t = 120 = 0.8T$ (bottom), in grey and in the form of probability densities. The black curves indicate the PDF of the respective statistics obtained via a Monte-Carlo simulation involving independent standard normal random variables. Note that the vertical axis is logarithmic

is included in the definition of RMSE, which makes it like a deterministic verification metric since the ensemble mean can be verified as if it were a deterministic prediction in its own right. The fact that the CRPS and the RMSE can be functionally related provides a link between deterministic and probabilistic verification. Assuming that the predictive and verifying distributions are both normal, the relationship comes in the form

$$\text{CRPS}^* = \frac{f(b, r)}{\sqrt{1 + b^2 + r^2}} \frac{1}{\sqrt{1 + h}} \text{RMSE}^*, \quad (29)$$

where b is a measure of bias (Equation (7)), r is a measure of non-calibrated ensemble dispersion (Equation (8)), $f(b, r)$ is as given in Equation (10), and h is the relative heteroscedasticity of the ensemble's standard deviation σ_P as defined in Equation (21). The asterisks accompanying the notations CRPS and RMSE refer to aggregation over a sample, and the heteroscedasticity refers to the variability of σ_P across the dimensions of aggregation. The CRPS–RMSE relationship is subject to a technical assumption that the measures of non-calibration (b and r) do not depend on the predictive distribution.

When predictions are reliable, Equation (29) reduces to

$$\text{CRPS}^* = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{1 + h}} \text{RMSE}^*. \quad (30)$$

The relationship in this special case has been tested on simulations of idealised 2D turbulence (Section 4), in which ensembles are reliable by the experimental design. Heteroscedastic effects are present, and are found to depend considerably on the length-scale and the forecast lead time. To our knowledge, the origins of such heteroscedastic effects in idealised turbulence are not well understood. It would be interesting to investigate the scale-dependence of heteroscedasticity in the future. In any case, if we were to ignore such heteroscedastic effects, a simpler form of the CRPS–RMSE relationship

$$\text{CRPS}^* = \frac{1}{\sqrt{2\pi}} \text{RMSE}^* \quad (31)$$

would hold, thus making the CRPS a constant multiple of the RMSE. Equation (31) turns out to be a reasonably good approximation of the CRPS–RMSE relationship recorded in the numerical simulations, including the times at which the two error metrics saturate at different scales. Deviations from this equation remain within a factor of two. When heteroscedasticity is taken into account, the two sides of Equation (30) agree with excellent accuracy. On the other hand, our results show that the CRPS–RMSE relationship is resilient to non-normality in ensemble distributions, at least to the extent demonstrated by this experiment. Moreover, the factor $1/\sqrt{2\pi}$ in Equation (31) is robust to small ensemble biases, although it is more sensitive to over- and underdispersion of the ensemble.

The CRPS–RMSE relationship may be applied on any scalar meteorological variable in the real world, provided that the distribution of the variable is not overly non-normal. Inhomogeneity and anisotropy of the atmospheric flow imply that the results will depend on the domain and direction of aggregation. It remains to be seen how the heteroscedasticity observed in NWP simulations compares with that reported here for idealised 2D turbulence.

AUTHOR CONTRIBUTIONS

Tsz Yan Leung: conceptualization; formal Analysis; writing Original Draft. **Martin Leutbecher:** methodology; resources; supervision; writing Review Editing. **Sebastian Reich:** funding Acquisition; supervision; writing Review Editing. **Theodore G. Shepherd:** funding Acquisition; supervision; writing Review Editing.

ACKNOWLEDGEMENTS

Tsz Yan Leung was supported through a PhD scholarship awarded by the Engineering and Physical Sciences Research Council Grant EP/L016613/1 'EPSRC Centre for


Doctoral Training in the Mathematics of Planet Earth at Imperial College London and the University of Reading', with additional funding support from the European Research Council Advanced Grant 'Understanding the Atmospheric Circulation Response to Climate Change' (ACRCC), project 339390, under Theodore G. Shepherd as the Principal Investigator. The work of Sebastian Reich has been partially funded by Deutsche Forschungsgemeinschaft (DFG, German Science Foundation) – SFB 1114/2 235221301. The authors thank Richard Scott for providing his code for modification for the numerical simulations in Section 4, which would have not been possible without further support from high-performance computing resources at the European Centre for Medium-Range Weather Forecasts. The authors also wish to thank the two anonymous reviewers for their helpful and valuable comments on an earlier version of the manuscript.

AUTHORS' DECLARATION

An earlier and expanded version of this article forms part of the first author's PhD thesis (Leung, 2020).

ORCID

Tsz Yan Leung  <https://orcid.org/0000-0003-0056-284X>

Martin Leutbecher  <https://orcid.org/0000-0003-4160-0750>

REFERENCES

- Baringhaus, L. and Franz, C. (2004) On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88, 190–206.
- Bauer, P., Thorpe, A. and Brunet, G. (2015) The quiet revolution of numerical weather prediction. *Nature*, 525, 47–55.
- Farge, M., Schneider, K. and Kevlahan, N. (1999) Non-Gaussianity and coherent vortex simulation for two-dimensional turbulence using an adaptive orthogonal wavelet basis. *Physics of Fluids*, 11, 2187–2201.
- Gneiting, T. and Raftery, A.E. (2007) Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102, 359–378.
- Hersbach, H. (2000) Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15, 559–570.
- Judt, F. (2020) Atmospheric predictability of the tropics, middle latitudes, and polar regions explored through global storm-resolving simulations. *Journal of the Atmospheric Sciences*, 77, 257–276.
- Leith, C.E. (1974) Theoretical skill of Monte Carlo forecasts. *Monthly Weather Review*, 102, 409–418.
- Leung, T.Y. (2020). Weather Predictability: Some Theoretical Considerations. PhD thesis, University of Reading, UK.
- Leung, T.Y., Leutbecher, M., Reich, S. and Shepherd, T.G. (2020) Impact of the mesoscale range on error growth and the limits to atmospheric predictability. *Journal of the Atmospheric Sciences*, 77, 3769–3779.
- Leutbecher, M. (2019) Ensemble size: how suboptimal is less than infinity?. *Quarterly Journal of the Royal Meteorological Society*, 145(Suppl. 1), 107–128.
- Leutbecher, M. and Haiden, T. (2021) Understanding changes of the continuous ranked probability score using a homogeneous Gaussian approximation. *Quarterly Journal of the Royal Meteorological Society*, 147, 425–442.
- Lorenz, E.N. (1969) The predictability of a flow which possesses many scales of motion. *Tellus*, 21, 289–307.
- Nastrom, G.D. and Gage, K.S. (1985) A climatology of atmospheric wavenumber spectra of wind and temperature observed by commercial aircraft. *Journal of the Atmospheric Sciences*, 42, 950–960.

How to cite this article: Leung, T.Y., Leutbecher, M., Reich, S. & Shepherd, T.G. (2021) Forecast verification: Relating deterministic and probabilistic metrics. *Quarterly Journal of the Royal Meteorological Society*, 147(739), 3124–3134. Available from: <https://doi.org/10.1002/qj.4120>

APPENDIX. EXPECTED CRPS FOR NORMAL PREDICTIVE AND VERIFYING DISTRIBUTIONS

The integral

$$\int_{-\infty}^{\infty} \sigma_Q \frac{r^2}{\sqrt{\pi}} \left\{ -1 + \sqrt{\pi} x \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right) + \sqrt{2} \exp\left(-\frac{1}{2}x^2\right) \right\} \varphi(rx + b) dx \quad (\text{A1})$$

in Equation (6) can be simplified to provide an analytic expression for CRPS(P, Q), the expected CRPS for normal predictive and verifying distributions. The integral will be decomposed into three contributions according to the terms inside the outermost parentheses of the integrand. These contributions will be evaluated one by one. To begin, we have

$$\int_{-\infty}^{\infty} \sigma_Q \frac{r^2}{\sqrt{\pi}} \{-\varphi(rx + b)\} dx = -\sigma_Q \frac{r}{\sqrt{\pi}} \quad (\text{A2})$$

and

$$\begin{aligned} & \int_{-\infty}^{\infty} \sigma_Q \frac{r^2}{\sqrt{\pi}} \sqrt{2} \exp\left(-\frac{1}{2}x^2\right) \varphi(rx + b) dx \\ &= \sigma_Q r^2 \sqrt{\frac{2}{\pi(1+r^2)}} \exp\left(-\frac{1}{2} \frac{b^2}{1+r^2}\right), \end{aligned} \quad (\text{A3})$$

since they are Gaussian integrals. As for the contribution

$$\int_{-\infty}^{\infty} \sigma_Q r^2 x \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right) \varphi(rx+b) dx, \quad (\text{A4})$$

we proceed by first seeking an indefinite integral $A(x)$ of $x\varphi(rx+b)$, so that Expression A4 can be written as

$$\sigma_Q r^2 \left[A(\cdot) \operatorname{erf}\left(\frac{\cdot}{\sqrt{2}}\right) \right]_{-\infty}^{\infty} - \sigma_Q r^2 \int_{-\infty}^{\infty} A(x) d \left\{ \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right) \right\}. \quad (\text{A5})$$

It is not difficult to see that

$$\begin{aligned} A(x) &= \int_{-\infty}^x x' \varphi(rx'+b) dx' \\ &= -\frac{1}{\sqrt{2\pi}r^2} \exp\left\{-\frac{1}{2}(rx+b)^2\right\} - \frac{b}{r^2} \Phi(rx+b), \quad (\text{A6}) \end{aligned}$$

so that $A(\infty) = -b/r^2$ and $A(-\infty) = 0$. Substituting these into (A5), the first term equals $-b\sigma_Q$, whereas

$$\begin{aligned} & -\sigma_Q r^2 \int_{-\infty}^{\infty} A(x) d \left\{ \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right) \right\} \\ &= \sigma_Q r^2 \int_{-\infty}^{\infty} \left[\frac{1}{\sqrt{2\pi}r^2} \exp\left\{-\frac{1}{2}(rx+b)^2\right\} + \frac{b}{r^2} \Phi(rx+b) \right] \\ & \quad \times d\{2\Phi(x) - 1\} \\ &= \sqrt{\frac{2}{\pi}} \sigma_Q \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}(rx+b)^2\right\} \varphi(x) dx \\ & \quad + 2b\sigma_Q \int_{-\infty}^{\infty} \Phi(rx+b) \varphi(x) dx. \quad (\text{A7}) \end{aligned}$$

The first term on the right-hand-side of Equation (A7) is a Gaussian integral which evaluates to

$$\sigma_Q \sqrt{\frac{2}{\pi(1+r^2)}} \exp\left(-\frac{1}{2} \frac{b^2}{1+r^2}\right).$$

The second term equals

$$\begin{aligned} & b\sigma_Q \int_{-\infty}^{\infty} \left\{ 1 + \operatorname{erf}\left(\frac{rx+b}{\sqrt{2}}\right) \right\} \varphi(x) dx \\ &= b\sigma_Q \left\{ 1 + \operatorname{erf}\left(\frac{b}{\sqrt{2(1+r^2)}}\right) \right\}, \quad (\text{A8}) \end{aligned}$$

which can be established by considering

$$I(r,b) := \int_{-\infty}^{\infty} \operatorname{erf}\left(\frac{rx+b}{\sqrt{2}}\right) \varphi(x) dx \quad (\text{A9})$$

and writing

$$I(r,b) = \int_0^b \frac{\partial I}{\partial b}(r,b') db'$$

(note that $I(r,0) = 0$, as the integrand is in that case an odd function). Hence we can write (A4) as

$$\begin{aligned} & -b\sigma_Q + \sigma_Q \sqrt{\frac{2}{\pi(1+r^2)}} \exp\left(-\frac{1}{2} \frac{b^2}{1+r^2}\right) \\ & + b\sigma_Q \left\{ 1 + \operatorname{erf}\left(\frac{b}{\sqrt{2(1+r^2)}}\right) \right\} \\ &= \sigma_Q \sqrt{\frac{2}{\pi(1+r^2)}} \exp\left(-\frac{1}{2} \frac{b^2}{1+r^2}\right) \\ & + b\sigma_Q \operatorname{erf}\left(\frac{b}{\sqrt{2(1+r^2)}}\right) \quad (\text{A10}) \end{aligned}$$

Substituting this and Equations (A2) and (A3) into (A1) and therefore Equation (6), we finally arrive at

$$\text{CRPS}(P,Q) = \sigma_Q f(b,r), \quad (\text{A11})$$

where

$$\begin{aligned} f(b,r) &= -\frac{r}{\sqrt{\pi}} + \sqrt{\frac{2(1+r^2)}{\pi}} \exp\left(-\frac{b^2}{2(1+r^2)}\right) \\ & + b \operatorname{erf}\left(\frac{b}{\sqrt{2(1+r^2)}}\right). \quad (\text{A12}) \end{aligned}$$