The Continuous Ranked Probability Score for Circular Variables and its Application to Mesoscale Forecast Ensemble Verification

Eric P. Grimit[a], Tilmann Gneiting[b], Veronica Berrocal[b] and Nicholas A. Johnson[c]

Department of Atmospheric Sciences[a] and Department of Statistics[b]
University of Washington

Department of Statistics, Stanford University[c]

Technical Report no. 493

Department of Statistics, University of Washington

January 2006

Abstract

An analogue of the linear continuous ranked probability score is introduced that applies to probabilistic forecasts of circular quantities. This scoring rule is proper and thereby discourages hedging. The circular continuous ranked probability score reduces to angular distance when the forecast is deterministic, just as the linear continuous ranked probability score generalizes the absolute error. Furthermore, the continuous ranked probability score provides a direct way of comparing deterministic forecasts, discrete forecast ensembles, and post-processed forecast ensembles that can take the form of probability density functions.

The circular continuous ranked probability score is used in this study to assess predictions of 10 m wind direction for 361 cases of mesoscale, short-range ensemble forecasts over the North American Pacific Northwest. Reference probability forecasts based on the ensemble mean and its forecast error history over the period outperform probability forecasts constructed directly from the ensemble sample statistics. These results suggest that short-term forecast uncertainty is not yet well predicted at mesoscale resolutions near the surface, despite the inclusion of multi-scheme physics diversity and surface boundary parameter perturbations in the mesoscale ensemble design.[1]

Keywords: Proper scoring rule; von Mises distribution; Wind direction.

[1]This research was supported by the DoD Multidisciplinary University Research Initiative (MURI) program administered by the Office of Naval Research under Grant N00014-01-10745, and by the National Science Foundation under Award 0134264. In Academic Year 2004/05, Tilmann Gneiting was on sabbatical leave at the Soil Physics Group, Universität Bayreuth, Universitätsstr. 30, 95440 Bayreuth, Germany.

1 Introduction

Scoring rules can be used to assess the quality of weather forecasts by assigning a numerical score based on the forecast and the value or event that materializes. A *proper* scoring rule maximizes the expected reward (or minimizes the expected penalty) for forecasting one's true beliefs, thereby discouraging hedging or cheating (Jolliffe and Stephenson 2003, pp. 8, 27; Wilks 2006, p. 298).

In evaluating probabilistic weather forecasts of linear, real-valued variables, such as temperature and pressure, the *continuous ranked probability score* (CRPS; Matheson and Winkler 1976; Unger 1985) has attracted renewed attention by the meteorological community (Hersbach 2000; Candille and Talagrand 2005; Gneiting *et al.* 2005; Wilks 2006, p. 302). If F is the cumulative distribution function of the forecast distribution and x verifies, the CRPS is defined as

$$\mathrm{CRPS}(F, x) = \int_{-\infty}^{\infty} \left(F(y) - \mathbf{1}\{y \geq x\}\right)^2 \, \mathrm{d}y, \tag{1}$$

where $\mathbf{1}\{y \geq x\}$ denotes a step function along the real line that attains the value 1 if $y \geq x$ and the value 0 otherwise. Gneiting and Raftery (2004) show that the continuous ranked probability score can be written equivalently as

$$\mathrm{CRPS}(F, x) = \mathrm{E}\{\, |X - x| \,\} - \frac{1}{2}\,\mathrm{E}\{\, |X - X^*| \,\}, \tag{2}$$

where X and X^* are independent copies of a linear random variable with distribution function F, and $\mathrm{E}\{\cdot\}$ denotes the expectation operator. The CRPS is proper and is expressed in the same unit as the observed variable. The CRPS generalizes the absolute error, and therefore provides a direct way of comparing various deterministic and probabilistic forecasts using a single metric.

For several reasons, near-surface wind direction is an important meteorological variable to consider when undertaking mesoscale forecast verification studies. Wind direction mitigates some of the problems caused by low observation density at mesoscale resolutions because wind direction is inherently linked with the local pressure gradient. Hence, wind direction observations yield information about the atmospheric state in a region around the station location, both horizontally and through depth. In addition, routine surface wind observing sites are typically more numerous than those with pressure sensors. Effectively, the use of wind direction increases the density of the pressure observing network while avoiding potential problems with below-ground pressure reduction assumptions and calculation of spatial derivatives.

If mesoscale perturbations to the large-scale flow are present, either as a result of strong synoptic-scale flow impinging on topography (e.g., lee troughs) or because of differential heating/cooling between adjacent regions under weak synoptic-scale forcing (e.g., land-sea breeze circulations), routine surface wind direction observations provide a means of detecting their presence (but not necessarily their strength). Validation of surface wind direction predictions addresses a central purpose of mesoscale numerical weather prediction models:

their ability to simulate important local circulations. A long-term mesoscale forecast verification study over the North American Pacific Northwest shows that 10 m wind direction statistics using mean absolute errors have value in determining the impact of increased horizontal resolution on deterministic forecast accuracy (Mass *et al.* 2002).

To extend this strategy to mesoscale forecast ensemble verification, a tool is needed that assesses probabilistic wind direction forecasts. Wind direction is a circular variable, and therefore standard scoring rules for verifying probabilistic forecasts of linear variables do not apply. In particular, the representations (1) and (2) of the *linear* CRPS are based on Euclidean distance measures that are inappropriate for directional variables. To address these issues, a *circular* analogue of the linear CRPS is introduced, defined by

$$\mathrm{CRPS}_{\mathrm{circ}}(P, \theta) = \mathrm{E}\{\alpha(\Theta, \theta)\} - \frac{1}{2}\,\mathrm{E}\{\alpha(\Theta, \Theta^*)\},$$

where P is a forecast distribution on the circle, θ is the verifying direction, α is angular distance, and Θ and Θ^* are independent copies of a circular random variable with distribution P. This replaces the Euclidean distance in the representation (2) of the linear CRPS by the natural distance metric for directional variables: the angular distance. The circular CRPS is proper and reduces to angular distance when the forecast is deterministic, just as the linear CRPS generalizes the absolute error. It is reported in units of angular distance and allows for the direct comparison of deterministic forecasts, discrete forecast ensembles, and post-processed forecast ensembles that can take the form of a probability density function.

The remainder of the paper is organized as follows. Section 2 provides a review of the linear CRPS and introduces its circular analogue. The proof of the propriety of the circular CRPS is deferred to the Appendix. Sections 3 and 4 illustrate the use of the circular CRPS in assessing predictions of 10 m wind direction for 361 cases of mesoscale, short-range ensemble forecasts over the North American Pacific Northwest. Reference probability forecasts of wind direction, constructed from the observed history of ensemble-mean forecast errors, are found to produce lower average circular CRPS than probability forecasts based on the direct short-range ensemble output. The paper closes with a discussion of the implications of these verification results in Section 5.

2 The continuous ranked probability score (CRPS)

This section contains a review of the properties of the linear CRPS and an introduction to its circular analogue.

2.1 The CRPS for linear variables

The *linear* CRPS is defined in terms of the forecast cumulative distribution function F and the verifying observation x of a linear, real-valued variable as in (1). Traditionally, this score is interpreted as the integral of the Brier score for binary probability forecasts at all real-valued thresholds y (Hersbach 2000; Toth *et al.* 2003). The alternative, yet equivalent,

representation given by (2) makes the interpretation of CRPS lucid. The first term on the right-hand side of (2) is the expected value of the absolute error, and the second term is a correction factor that measures the sharpness of the probabilistic forecast F and renders the score proper. The linear CRPS generalizes the absolute error, to which it reduces if F is a deterministic forecast. If $F = F_{\text{ens}}$ is a discrete predictive distribution from a forecast ensemble of size M, the evaluation of the CRPS is straightforward. The predictive cumulative distribution function F_{ens} has jumps of size $1/M$ at the ensemble member values $x_1, \ldots, x_M$, and (2) reduces to

$$\text{CRPS}(F_{\text{ens}}, x) = \frac{1}{M} \sum_{m=1}^{M} |x_m - x| - \frac{1}{2M^2} \sum_{m=1}^{M} \sum_{n=1}^{M} |x_m - x_n|. \tag{3}$$

Thus, the linear CRPS provides a direct way of comparing deterministic and probabilistic forecasts with a single metric that is proper and reported in the same unit as the observations, making it a tangible measure of predictive performance. Additionally, Hersbach (2000) and Candille and Talagrand (2005) describe reliability-resolution (or calibration-sharpness) decompositions of the linear CRPS.

The linear CRPS and the negative of the *logarithmic score* (*information deficit* or *ignorance score*) are seeing increased use in probabilistic weather forecast evaluations (Unger 1985; Hersbach 2000; Roulston and Smith 2002; Grimit 2004; Candille and Talagrand 2005; Gneiting *et al.* 2005). Although the logarithmic score is proper and trivial to calculate, it involves a harsh penalty for low probability events and therefore is highly sensitive to outliers (Selten 1998). Furthermore, the logarithmic score cannot be used to assess deterministic or discrete ensemble forecasts. The *spherical score* and the *quadratic score* (Matheson and Winkler 1976) share this constraint. The linear CRPS provides a more resistant and more flexible alternative.

The lack of analytic solutions to the defining integral in (1) has restrained widespread use of the linear CRPS as a scoring rule. Recently, Gneiting *et al.* (2005) derived an analytic expression for the linear CRPS when the forecast distribution function F is Gaussian with mean μ and variance σ^2. Using (2), they showed that

$$\text{CRPS}(\mathcal{N}(\mu, \sigma^2), x) = \sigma\biggl(\frac{x-\mu}{\sigma} \biggl(2\Phi\biggl(\frac{x-\mu}{\sigma} \biggr) - 1 \biggr) + 2\phi\biggl(\frac{x-\mu}{\sigma} \biggr) - \frac{1}{\sqrt{\pi}} \biggr), \tag{4}$$

where ϕ and Φ represent the standard Gaussian probability density and cumulative distribution functions, respectively. Similarly, closed form solutions can be found for other forecast distributions, including mixtures of Gaussian distributions, which are used in the Bayesian model averaging approach to ensemble calibration (Raftery *et al.* 2005). Using (2), the CRPS for a Gaussian mixture distribution is

$$\begin{aligned} \text{CRPS}\biggl(\sum_{m=1}^{M} w_m \mathcal{N}(\mu_m, \sigma_m^2), x \biggr) = \\ \sum_{m=1}^{M} w_m A\Bigl(x - \mu_m, \sigma_m^2 \Bigr) - \frac{1}{2} \sum_{m=1}^{M} \sum_{n=1}^{M} w_m w_n A\Bigl(\mu_m - \mu_n, \sigma_m^2 + \sigma_n^2 \Bigr). \end{aligned} \tag{5}$$

In this equation, w_i is the weight for ensemble member i, with the weights being nonnegative and summing to 1, and

$$A\Big(\mu, \sigma^2\Big) = 2\sigma\,\phi\bigg(\frac{\mu}{\sigma}\bigg) + \mu\,\bigg(2\,\Phi\bigg(\frac{\mu}{\sigma}\bigg) - 1\bigg) \tag{6}$$

denotes the expectation of the absolute value of a normal random variable with mean μ and variance σ^2. Gneiting *et al.* (2006) give an analytic expression for the linear CRPS when a Gaussian forecast distribution is truncated at zero, thereby taking account of the potential nonnegativity of a variable such as wind speed.

2.2 The CRPS for circular variables

Any analogue of the linear CRPS for circular variables needs to be proper. In addition, the score should be independent of the verifying direction when the forecast distribution is uniform on the circle. The direct analogue of (1) does not satisfy these requirements. Hence, the circular CRPS should follow the representation (2) and be expressed as

$$\mathrm{CRPS}_{\mathrm{circ}}(P, \theta) = \mathrm{E}\{\,\alpha(\Theta, \theta)\,\} - \frac{1}{2}\,\mathrm{E}\{\,\alpha(\Theta, \Theta^*)\,\}, \tag{7}$$

where $\alpha(\theta, \theta^*) \in [0, \pi]$ denotes the angular distance between any two directions θ and θ^* on the circle $[-\pi, \pi)$, and where Θ and Θ^* are independent random variables with common circular probability distribution P. In the Appendix, Fourier analytic tools are used to prove the propriety of the circular CRPS. Note that, if P is uniform on the circle then (7) reduces to $\pi/4$ and is independent of the verifying direction.

The circular CRPS reduces to angular distance when the forecast is deterministic, just as the linear CRPS generalizes the absolute error. For a discrete predictive distribution from a forecast ensemble $\theta_1, \ldots, \theta_M$, its evaluation is straightforward. The circular predictive distribution P_{ens} has probability mass $1/M$ at $\theta_1, \ldots, \theta_M$, and (7) becomes

$$\mathrm{CRPS}_{\mathrm{circ}}(P_{\mathrm{ens}}, \theta) = \frac{1}{M}\sum_{m=1}^{M} \alpha(\theta_m, \theta) - \frac{1}{2M^2}\sum_{m=1}^{M}\sum_{n=1}^{M} \alpha(\theta_m, \theta_n). \tag{8}$$

The circular CRPS is reported in units of angular distance and provides a direct way of comparing deterministic forecasts, discrete forecast ensembles, and post-processed forecast ensembles that can take the form of probability density functions from a circular parametric family.

The von Mises distribution is the most common parametric model for unimodal samples of circular data (Fisher 1993). In many respects, it forms the natural, circular analogue of the Gaussian distribution on the real line. The probability density function of the von Mises distribution with mean direction $\mu \in [-\pi, \pi)$ and concentration parameter $\kappa \geq 0$ is

$$f_{\mathrm{VM}}(\theta) = [2\pi\, I_0(\kappa)]^{-1} \exp[\kappa \cos(\theta - \mu)], \qquad -\pi \leq \theta < \pi\,, \tag{9}$$

where I_0 denotes the modified Bessel function of the first kind of order zero. As κ tends to zero, the von Mises distribution converges to the uniform distribution on the circle. As κ approaches infinity, the von Mises distribution becomes increasingly concentrated and can be approximated by a Gaussian distribution with mean μ and variance $1/\kappa$. A sample estimate of κ can be obtained with the maximum-likelihood technique using the algorithm of Best and Fisher (1981).

If a probabilistic forecast takes the form of a von Mises distribution, the first term on the right-hand side of (7) can be written as an integral,

$$\mathrm{E}\{\,\alpha(\Theta,\theta)\,\} = [2\pi I_0(\kappa)]^{-1} \int_{-\pi}^{\pi} \alpha(y,\theta) \exp[\kappa \cos(y-\mu)]\,\mathrm{d}y, \tag{10}$$

which is readily evaluated by standard numerical integration routines. As κ tends to infinity, the aforementioned Gaussian approximation implies that (10) is asymptotically equivalent to $A(\alpha(\mu,\theta), 1/\kappa)$, where the function A is defined in (6). We use this approximation if $\kappa \geq 1500$. The second term on the right-hand side of (7) is a function of the concentration parameter κ only, say $g(\kappa)$. If $\kappa = 0$ the von Mises distribution reduces to the uniform distribution on the circle; hence $g(0) = \pi/4$. As κ tends to infinity, the above argument shows that $g(\kappa)$ is asymptotically equivalent to $1/(2\pi\kappa)^{1/2}$. We apply this approximation if $\kappa \geq 200$. If $\kappa < 200$, we use standard Monte Carlo integration to compute estimates of $g(\kappa)$, and we apply the lowess technique (Cleveland 1979) to smooth the resulting lookup-table. This provides an algorithm that approximates $g(\kappa)$, $\kappa \in [0,\infty)$, to at least three decimal places. Code is available at `cran.r-project.org` in the form of the function `crps.circ` within the R verification package (Pocernich 2006). FORTRAN code can also be made available on request.

3 Mesoscale forecast ensemble data

The mesoscale ensemble prediction system, the post-processing techniques applied to the output, and the verification data used in this study are now described.

3.1 The University of Washington short-range ensemble forecast (UW SREF) systems

Over the last few years, a short-range ensemble prediction system for the North American Pacific Northwest has been under development, with the goal of producing calibrated forecast probabilities for near-surface weather parameters at mesoscale resolutions. The original five-member mesoscale ensemble was designed as a single-model, multi-analysis ensemble prediction system using MM5 with a nested, limited-area grid configuration focusing on Washington and Oregon (Grimit and Mass 2002). The horizontal grid spacing of the inner nest was 12 km and 33 sigma levels were used in the vertical. Large-scale analyses and forecasts from several operational forecast centers provided the necessary initial conditions (ICs) and lateral boundary conditions (LBCs). Beginning in the autumn of 2002, the size

of the mesoscale ensemble was increased to eight members using additional global analyses and forecasts and named the University of Washington Mesoscale Ensemble (UWME; Eckel and Mass 2005).

Although UWME produces skillful forecast probabilities without any post-processing, the forecast dispersion is inadequate. Eckel and Mass (2005) attribute a primary cause of low dispersion to the lack of model error representation within the UWME system design. To test whether model physics diversity boosts ensemble dispersion in an appropriate manner, a parallel eight-member system (UWME+) is run with the same configuration of ICs and LBCs, but each member uses a different combination of MM5 physics options for planetary boundary layer, cloud microphysics, cumulus convection, and radiation parametrizations. Surface parameter uncertainty is also incorporated into UWME+ by perturbing the sea-surface temperature lower-boundary condition and the fixed parameters describing soil-moisture availability, albedo, and surface roughness length. Inclusion of model physics diversity and surface parameter uncertainty increases forecast dispersion toward statistical consistency and thus improved mesoscale forecast probability skill. A complete description of UWME+ and its design is contained in Eckel and Mass (2005). These UW SREF systems were run to 48 h lead time beginning at 0000 UTC each day.

3.2 Verification data and evaluation period

The evaluation period of this study begins on 31 October 2002 and extends through 31 March 2004. This period encompasses two cool seasons (October–March) during 2002–03 and 2003–04 and one warm season (April–September) during 2003. Out of the 518 total days, 361 days are identified where all UWME and UWME+ member forecasts are complete and verification data is available. Of the 361 total cases, 271 are cool season cases and 90 are warm season cases. Verification statistics are compiled separately for each forecast lead time at three hour intervals through 48 h, since model performance varies substantially throughout the diurnal cycle and generally degrades as the lead time increases. Circular CRPS is computed between forecast-observation pairs only when the verifying wind speed is at least 5 knots (2.57 m/s), since wind direction observations are unreliable at lower wind speeds.

Verification data includes the use of both station-based observations and gridded surface analyses. The operational 20-km Rapid Update Cycle (RUC20; Benjamin *et al.* 2004) mesoscale analysis from NCEP is used to represent truth for grid-based verification of the forecast ensembles. To make the comparison at a common horizontal resolution, the 12 km UW SREF output is fit to the native RUC20 grid using bi-linear interpolation. The 10 m wind components are subsequently rotated to account for the new center longitude. Only RUC20 grid boxes lying within the UW SREF 12 km domain and its five grid-box wide boundary region are considered. This yields 3490 common grid boxes for verification.

Station-based observations of near-surface wind direction are acquired in real-time from over two dozen networks operated by local, state, federal, and foreign agencies as well as a few independent organizations (Mass *et al.* 2003). All together, the mesoscale observation network includes approximately 800 wind observing locations within the UW SREF 12 km

domain. Model 10 m wind component forecasts at the four grid-box centers surrounding each station are bi-linearly interpolated to the observation location and then rotated from grid-relative to north-relative. No adjustment is made for any vertical displacement of the model surface level from the real terrain.

3.3 UW SREF post-processing

Several post-processing methods are applied to the wind direction forecast ensemble data and then compared using the circular CRPS. First, a bias correction is applied to each ensemble-member forecast separately, using bias estimates obtained from a simple, 14 day running-mean of the forecast errors at each grid box or station location. Eckel and Mass (2005) show that such a bias correction substantially improves both reliability and resolution of the forecast probabilities and also facilitates a fair comparison of methods, since systematic bias is not a component of forecast uncertainty. Second, an ensemble-mean forecast is generated by treating the individual ensemble-member forecasts as unit vectors in polar coordinates, summing them, and then finding the direction of the resultant vector sum (Fisher 1993, p. 31). Third, a continuous probability density forecast is generated from the bias-corrected forecast ensemble by fitting a von Mises distribution with the mean parameter given by the ensemble mean and the concentration parameter estimated with the Best and Fisher (1981) algorithm. Fourth, a reference probability density forecast is generated by dressing the bias-corrected ensemble mean using a von Mises distribution with a concentration parameter that is estimated from the ensemble-mean angular forecast errors realized over the entire sample period. Such a forecast has a static circular variance at each unique location and is an approximation to the climatological forecast uncertainty for the ensemble-mean wind direction. Since this forecast probability density is always centered on the bias-corrected ensemble mean, the mean remains state-dependent.

4 Results

To illustrate the calculation of circular CRPS for each post-processed forecast type, specific examples are taken from 24 h UWME forecasts of wind direction at Medford, Oregon (KMFR; 42.38°N, 122.87°W) and Victoria, British Columbia (CYYJ; 48.65°N, 123.43°W) valid on 0000 UTC 31 October 2003 (Fig. 1). Each panel shows a forecast probability mass function or probability density function on the unit circle corresponding to the ensemble mean (MEAN), the discrete (raw) forecast ensemble (RAW), the von Mises fit to the discrete forecast ensemble (FIT), and the von Mises fit to the MEAN error climatology (MEC) that is observed over the sample period.

The MEC forecast at KMFR (Fig. 1(a)) has a much lower von Mises concentration parameter than the FIT forecast, making MEC less sharp[2]. However, despite its lack of sharpness, MEC has lower (better) circular CRPS than FIT for this case because FIT is

[2]We expect MEC forecasts to be less sharp than FIT forecasts on average, since this is a defining characteristic of underdispersed ensemble prediction systems.

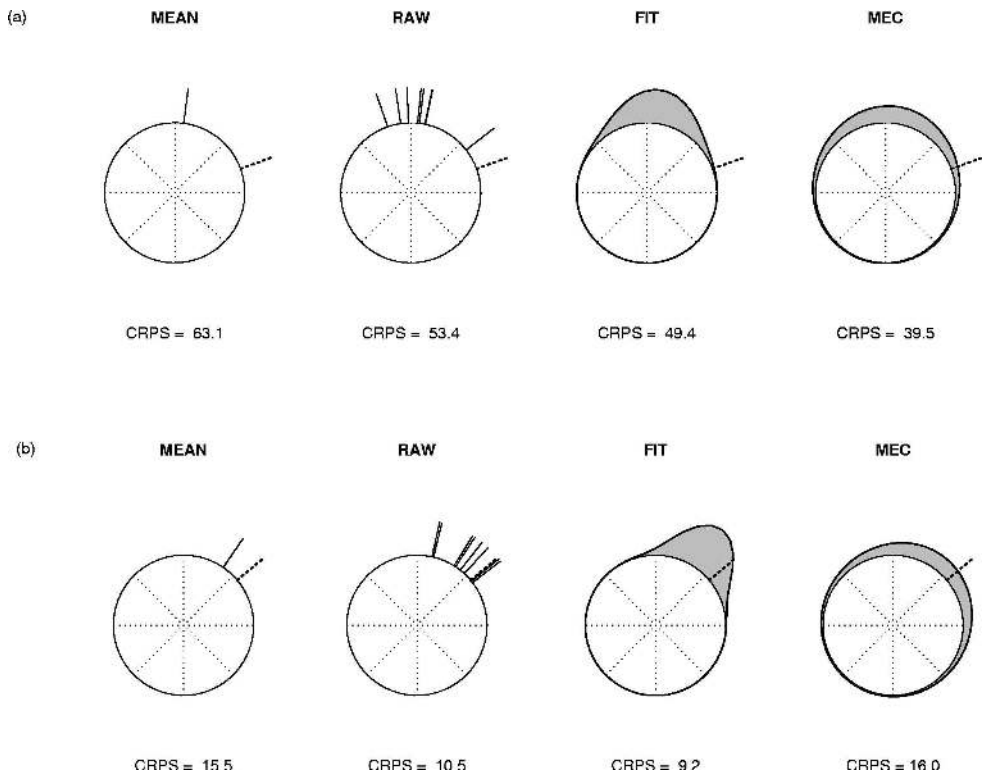

Figure 1: Circular diagrams of probability mass and probability density functions for 24 h UWME forecasts of wind direction (solid) at (a) Medford, Oregon (KMFR; 42.38°N, 122.87°W) and (b) Victoria, British Columbia (CYYJ; 48.65°N, 123.43°W) valid on 0000 UTC 31 October 2003. The forecasts correspond to the ensemble mean (MEAN), the discrete (raw) ensemble (RAW), the von Mises fit to the discrete ensemble (FIT), and the von Mises fit to the MEAN error climatology (MEC) over the sample period (in order, left to right). See text for details. Observed wind directions are dashed and the forecasts are solid. The actual circular CRPS (in degrees) is reported in each panel.

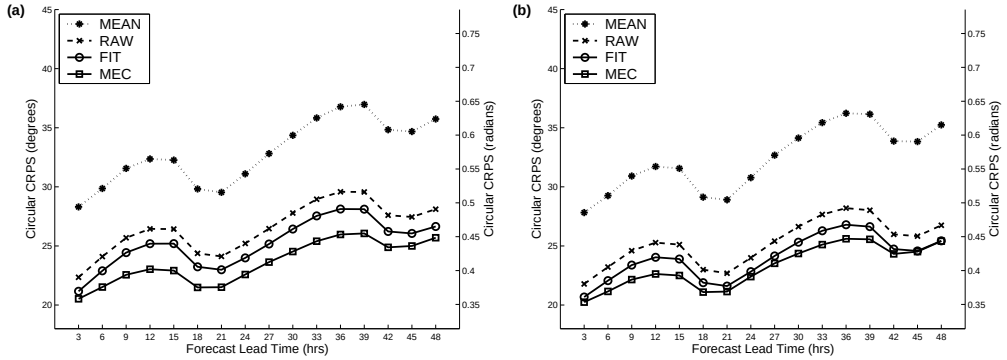

Figure 2: Spatio-temporal average circular CRPS by forecast lead time for 361 forecasts of 10 m wind direction over the North American Pacific Northwest during the period from October 2002 through March 2004 for the MEAN, RAW, FIT, and MEC forecasts produced from (a) the UWME system and (b) the UWME+ system. The 12 km UW forecasts are bi-linearly interpolated to the 20-km NCEP RUC (RUC20) mesoscale analyses that are used as truth and presented here on the portion of the RUC20 analysis grid that overlaps with the 12 km UW SREF grid.

more heavily penalized by the first term in (7), due to the large MEAN angular error. When the MEAN angular error is small, as it is for the forecast at CYYJ (Fig. 1(b)), FIT scores better than MEC as a direct result of the sharpness in the forecast ensemble.

Temporal averages of circular CRPS over the 361-case sample in the study period are calculated separately at each unique location and each forecast lead time. These average scores give an indication of the overall performance of the various 10 m wind direction forecasts from the 12 km UW SREF systems. When using RUC20 analyses as truth, independent statistics are saved at each NCEP analysis grid box that overlaps with the 12 km UW SREF grid. With station-based observations, CRPS statistics are unique to each point location. A summary of the 10 m wind direction forecast performance and an overall comparison of each ensemble system and its associated post-processing is conveyed by the spatial averages of these circular CRPS statistics (Figs. 2-3). In general, the UWME+ system performs better than the UWME system with slightly lower circular CRPS for all forecast types and lead times.

A large decrease in circular CRPS is observed between the MEAN forecasts and the RAW and FIT forecasts (which are very close to each other in performance). Even though all three forecast types are generated from the same raw ensemble values, circular CRPS decreases as one considers forecasts that are increasingly continuous in nature. The MEAN forecast is determinisitic, with all forecast probability mass placed at a single value. The RAW forecast is multi-valued, with disrete probability masses at several directions. The FIT forecast is a continuous probability density function. Evidently, the price paid in decreased sharpness is more than offset by the benefit of increased calibration. And since the CRPS is a proper scoring rule, the successive improvements cannot be considered a result of hedging.

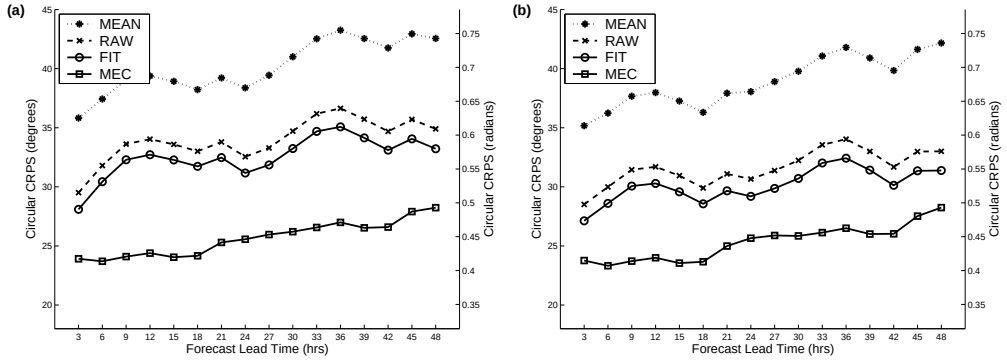

Figure 3: The same as in Fig. 2, except for 12 km ensemble forecasts bi-linearly interpolated to the surface observation sites that are used as truth.

For both ensemble systems and both verification methods, the MEC forecasts tend to perform the best. RAW and FIT forecasts from the UWME+ system do appear to be competitive with MEC at the latest forecast lead times, with FIT and MEC essentially having equal circular CRPS values at 45 h and 48 h. MEC forecasts have the advantage of being tuned by observations over a long period, resulting in nearly perfect calibration. However, such calibration comes at the expense of sharpness, and one would hope that the additional sharpness of FIT forecasts would lead to the best performance. Yet, these results show that MEC outperforms FIT on average. Even though the latter uncertainty forecasts have the advantage of being state-dependent and determined by the model dynamics, static uncertainty forecasts based on an approximation to the climatological errors of the ensemble mean perform better.

Observation-based verification yields larger circular CRPS for all forecast types than does grid-based verification. The differences in performance estimates are likely due to the disparate influence of representativeness errors[3] between the verification approaches. Another possibility could be that systematic errors in the RUC20 analyses lead to state estimates that are more similar to the MM5 model forecasts than to the real atmosphere.

Another discrepancy is the presence of a pronounced diurnal signal in the grid-based verification (Fig. 2) that is much weaker or even non-existent in the observation-based results (Fig. 3). The grid-based circular CRPS is higher for forecast lead times that correspond with overnight and early morning periods and lower during daytime hours. This characteristic is consistent with poor understanding of nocturnal boundary layer turbulence (Mahrt 1999) and the inability to model such regimes with existing atmospheric boundary layer parametrizations (e.g. Zhang and Zheng 2004). In particular, the MEC forecasts verified with observations do not show such a pattern, indicating that the average ensemble-mean error variance may oscillate little over the diurnal cycle. Thus, the results are inconclusive as to whether a diurnal cycle of 10 m wind direction errors is present. Representativeness errors may blur the diurnal cycle in observation-based verification and systematic errors in

[3]the forecast-observation scale mismatch

the RUC20 analyses, themselves diurnally varying, could artificially reduce the estimates of actual error.

Spatial maps of the circular CRPS illuminate what portions of the mesoscale domain experience relatively good or bad wind direction forecasts and where MEC forecasts make improvements over the others. Fig. 4 shows circular CRPS at 24 h for each of the four post-processed 10 m wind direction forecasts from the UWME+ system. Circular CRPS for MEAN forecasts, which is identical to the mean absolute error, ranges from below 20 degrees over the Pacific Ocean well offshore of Washington and Oregon on up to greater than 60 degrees over localized areas in eastern Washington and southern British Columbia (Fig. 4(a)). These areas of relatively larger 10 m wind direction forecast error seem to be located in low-lying areas adjacent to higher terrain, where the 12 km model may not sufficiently resolve narrow passes and valleys. Several additional pockets of circular CRPS greater than 40 degrees are apparent over many of the eastern slopes throughout the domain. This suggests that 10 m wind direction is also poorly forecast in the wake regions on the downstream side of ridges and mountains.

The circular CRPS for RAW, FIT, and MEC forecasts show spatial patterns that are very similar to the circular CRPS for the MEAN forecasts, but with reduced values nearly everywhere. Circular CRPS values less than 20 degrees now reach farther east, along the Oregon coast and into southwestern Washington. However, the percent improvement in circular CRPS for RAW, FIT, and MEC forecasts is small in areas where the MEAN circular CRPS is low. The largest improvements are realized over areas with the biggest circular CRPS. The circular CRPS for MEC forecasts ranges from well below 20 degrees to just over 40 degrees. As displayed in the examples from Fig. 1, MEC performs best where the MEAN forecast error is largest.

Spatial maps of the circular CRPS evaluated using station observations are contained in Fig. 5 and corroborate most of the results obtained through grid-based verification. The RAW and FIT forecasts sometimes worsen the circular CRPS at specific stations compared to the MEAN forecasts. However, MEC is nearly always the best and makes the most improvement at stations where the average MEAN circular CRPS is the largest.

Finally, it is noteworthy that many of the regions where the circular CRPS is largest in Fig. 4 have low observation density (Fig. 5). In fact, the areas in northeastern Washington and southern British Columbia that experience MEAN circular CRPS errors that average more than 60 degrees have no observations within them. Thus, it is likely that problems with the RUC20 analyses may contribute as much error as deficiencies in the numerical weather prediction model.

5 Discussion

In this paper, an analogue of the linear CRPS is introduced and extended to accommodate circular variables. The circular CRPS is proper and provides a simple way of comparing both deterministic and probabilistic circular forecasts with a single metric. Using this scoring rule, circular forecasts are rewarded for their sharpness and level of calibration. Predictions

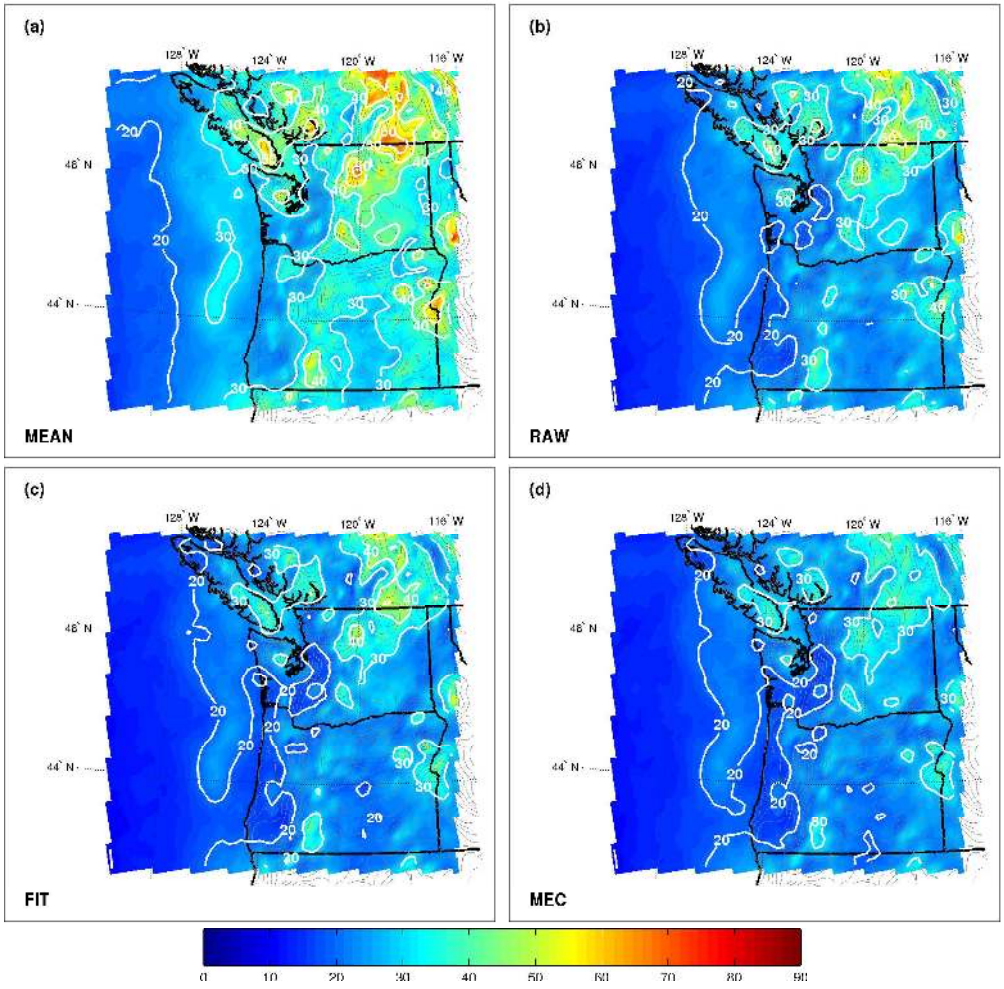

Figure 4: Spatial map of UWME+ circular CRPS (in degrees) at 24 h over RUC20 analysis grid boxes during the study period for (a) MEAN, (b) RAW, (c) FIT, and (d) MEC forecasts.

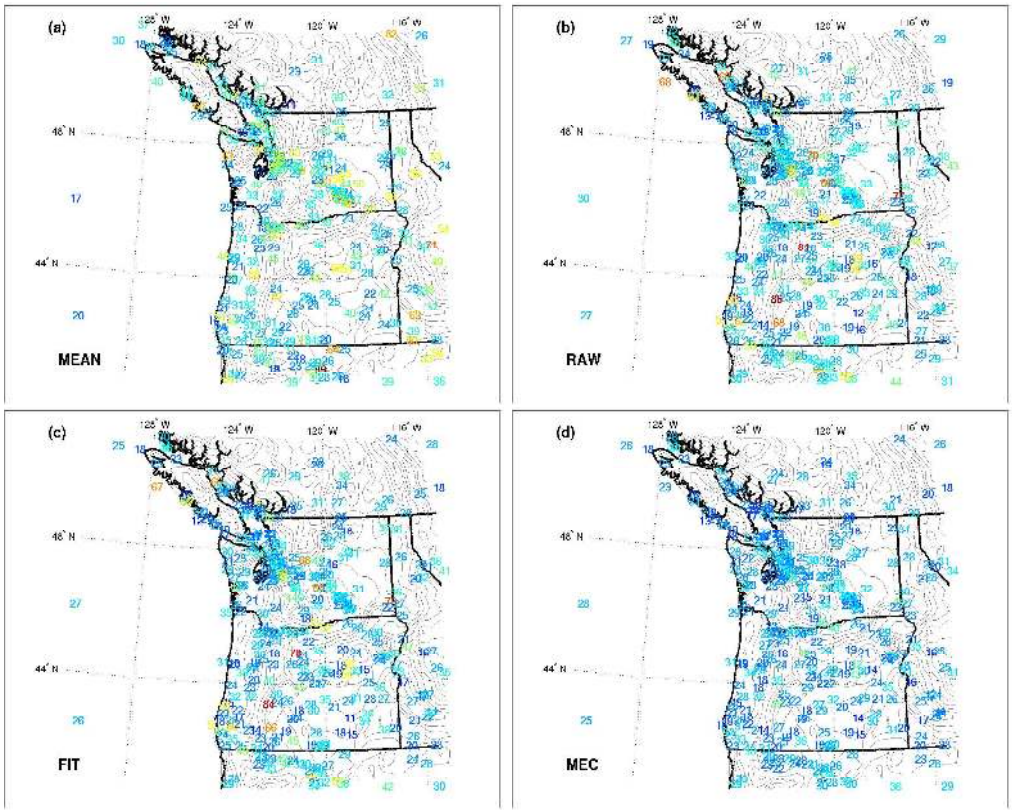

Figure 5: Spatial map of UWME+ circular CRPS (in degrees) at 24 h over surface observation sites during the study period for (a) MEAN, (b) RAW, (c) FIT, and (d) MEC forecasts. Color scale is the same as in Fig. 4.

of 10 m wind direction from the University of Washington short-range ensemble forecast systems are evaluated and compared over an 18-month period using this tool.

In general, the forecast ensemble system that incorporates multi-scheme physics diversity and surface parameter uncertainty into its design (UWME+) performs better than the ensemble system with initial condition diversity alone (UWME). The improvement in 10 m wind direction forecast skill realized by the UWME+ system appears to be less than that conveyed by Eckel and Mass (2005) for 2 m temperature and 10 m wind speed forecasts. A simple running-mean bias correction is applied in both ensemble verification studies and improves probability forecast skill.

Circular probability densities constructed directly from the forecast ensemble sample statistics (FIT) outperform the determinstic ensemble-mean (MEAN) 10 m wind direction predictions. However, a reference uncertainty forecast based on the sample climatological variance of ensemble-mean prediction errors (MEC) produces the lowest average circular CRPS. Such an improvement is realized because the MEC probability forecasts are calibrated, which more than offsets the degradation in circular CRPS that results from the reduced sharpness. MEC forecasts verify better (with lower circular CRPS) using both grid-based and observation-based verification data.

The superior performance of MEC forecasts suggests that short-term forecast uncertainty is not yet well predicted by the dynamic model ensembles at mesoscale resolutions near the surface. A static uncertainty forecast based on a local estimate of the climatological ensemble-mean prediction error variance produces better probabilistic forecasts of 10 m wind direction. Similar results have been obtained for 2 m temperature and 10 m wind speed forecasts from the UW SREF systems over the same period (Grimit 2004). Hence, it appears that several major sources of short-term mesoscale forecast error remain unaddressed by the ensemble prediction systems used in this study, despite the inclusion of multi-scheme physics diversity and surface parameter perturbations in the UWME+ system design. This suggests that more effort must be undertaken to incorporate additional sources of model uncertainty into the mesoscale ensemble prediction systems.

It is also apparent that more advanced statistical post-processing techniques must be applied to near-surface mesoscale ensemble predictions. Probability density forecasts ought to be calibrated and have better sharpness than MEC forecasts in order to possess skill. Several potential techniques have emerged toward meeting this challenge, including Bayesian model averaging (Raftery *et al.* 2005), two-step analogues from ensemble re-forecasts (Hamill *et al.* 2005), and dressing kernels with second moment constraints (Wang and Bishop 2005).

Proper scoring rules like CRPS provide composite measures of forecast performance that address calibration and sharpness simultaneously. Hersbach (2000) and Candille and Talagrand (2005) proposed calibration-sharpness (or reliability-resolution) decompositions of the linear CRPS. The practical computation of the decomposition can be strongly sensitive to implementation decisions (Candille and Talagrand 2005, p. 2144), and there is no obvious analogue of the decomposition that applies to the circular CRPS. However, calibration and sharpness can be assessed individually, using circular versions of the verification rank histogram and probability integral transform (PIT) histogram, respectively (Grimit

2001, pp. 70–71; Gneiting *et al.* 2005), and summary measures of the width of prediction intervals.

Single metrics of forecast performance like CRPS also do not take into account spatial or temporal displacements that may prove to be an important component of the true forecast quality for a particular user or application. Scoring rules are typically applied only to univariate forecasts at a fixed valid time, as in this study. Therefore, any inferences about forecast quality using only statistics based on such metrics should be judiciously applied.

Mesoscale forecast verification is a good example of a situation where caution should be used. Small displacements in space or time of otherwise well-predicted features can result in a double penalty if the forecast is scored at point locations with scalar metrics (e.g. Mass *et al.* 2002; Rife *et al.* 2004). The double-penalty effect can become more pronounced as grids with higher resolution are used and increasingly smaller wavelength features are resolved in the forecast. Hence, pointwise objective verification scores commonly do not show benefits that are commensurate with increased horizontal resolution. Some research efforts on alternative verification strategies have advocated the use of feature-based methods (Ebert and McBride 2000; Nachamkin 2004; Case *et al.* 2004; Done *et al.* 2004; Rife and Davis 2005). However, the density of routinely-taken observations near the surface is typically too low to verify small-wavelength features. The situation becomes especially problematic when features are interrupted or distorted by topography. Sufficient observational data is usually only available during field programs where a plethora of special observational platforms are deployed, including in situ aircraft, frequent radiosondes, ground-based radars, and additional surface observations (e.g. Bougeault *et al.* 2001; Stoelinga *et al.* 2003). However, such an effort is only sustainable for short periods and requires significant advance planning and financial resources.

Due to the relative sparsity of observations at mesoscale model resolutions, the question of how best to routinely verify these predictions remains an open one. The exclusive use of 10 m wind direction verification statistics does not fully solve the mesoscale verification problem. However, the issue of low observation density is at least partially addressed because the near-surface wind direction is strongly forced by the local horizontal pressure gradient, which is often well-resolved by the model grid spacing.

Appendix

In this appendix, the circular CRPS is shown to be a proper scoring rule that discourages hedging. The circle is identified with the interval $[-\pi, \pi)$. If θ and θ^* denote any two points on the circle, their angular distance is

$$\alpha(\theta, \theta^*) = \begin{cases} |\theta - \theta^*| & \text{if} \quad |\theta - \theta^*| \leq \pi, \\ 2\pi - |\theta - \theta^*| & \text{if} \quad \pi \leq |\theta - \theta^*| < 2\pi. \end{cases}$$

Let P be a probability distribution on the circle. Recall that, if the probabilistic forecast is P and $\theta \in [-\pi, \pi)$ verifies, the circular CRPS is defined as

$$\mathrm{CRPS}_{\mathrm{circ}}(P, \theta) = \mathrm{E}\{\alpha(\Theta, \theta)\} - \frac{1}{2}\mathrm{E}\{\alpha(\Theta, \Theta^*)\},$$

where Θ_P and Θ_P^* are independent copies of a circular random variable with distribution P, and E denotes the expectation operator.

Theorem. The circular continuous ranked probability score is a proper scoring rule. In other words,

$$\mathrm{E}\,\mathrm{CRPS}_{\mathrm{circ}}(P, \Theta_P) \leq \mathrm{E}\,\mathrm{CRPS}_{\mathrm{circ}}(Q, \Theta_P) \tag{11}$$

for all probability distributions P and Q on the circle.

Proof. Henceforth, Θ_P, Θ_P^* and Θ_Q, Θ_Q^* denote independent circular random variables with distributions P and Q, respectively. To prove the theorem, it suffices to show that

$$2\mathrm{E}\,\alpha(\Theta_P, \Theta_Q) - \mathrm{E}\,\alpha(\Theta_P, \Theta_P^*) - \mathrm{E}\,\alpha(\Theta_Q, \Theta_Q^*) \geq 0 \tag{12}$$

for all probability distributions P and Q on the circle.

Following Feller (1971, p. 633), the k-th *Fourier coefficient* of a circular distribution P is defined as the generally complex-valued number

$$p_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-ik\theta}\, P(\mathrm{d}\theta), \qquad k = 0, \pm 1, \pm 2, \ldots$$

In analogy to an argument of Székely (2003, p. 6), we find that

$$\begin{aligned} p_k\, \bar{q}_k + \bar{p}_k\, q_k &= \frac{1}{4\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \left(e^{ik(\theta - \theta^*)} + e^{-ik(\theta - \theta^*)} \right) P(\mathrm{d}\theta)\, Q(\mathrm{d}\theta^*) \\ &= \frac{1}{4\pi^2}\, \mathrm{E}\left(e^{ik(\Theta_P - \Theta_Q)} + e^{-ik(\Theta_P - \Theta_Q)} \right) \\ &= \frac{1}{2\pi^2}\, \mathrm{E} \cos(k\,(\Theta_P - \Theta_Q)), \end{aligned}$$

where p_k and q_k denote the Fourier coefficients of P and Q, respectively. Hence, we can write

$$\begin{aligned} |p_k - q_k|^2 &= p_k\, \bar{p}_k + q_k\, \bar{q}_k - (p_k\, \bar{q}_k + \bar{p}_k\, q_k) \\ &= \frac{1}{4\pi^2} \Big(\mathrm{E} \cos(k\,(\Theta_P - \Theta_P^*)) + \mathrm{E} \cos\Big(k \left(\Theta_Q - \Theta_Q^* \right) \Big) - 2\,\mathrm{E} \cos(k\,(\Theta_P - \Theta_Q)) \Big). \end{aligned}$$

Since k is an integer, the preceding equality, the definition of angular distance, and the cosine addition theorem imply that

$$\begin{aligned} 4\pi^2\, |p_k - q_k|^2 &= \\ &\mathrm{E} \left(\cos\left(k\, \alpha(\Theta_P, \Theta_P^*)\right) + \cos\left(k\, \alpha(\Theta_Q, \Theta_Q^*) \right) - 2 \cos\left(k\, \alpha(\Theta_P, \Theta_Q)\right) \right). \end{aligned} \tag{13}$$

Finally, we require equation (1.444.6) of Gradshteyn and Rhyzik (1994), which states that

$$\sum_{k=1}^{\infty} \frac{\cos\left((2k-1)\,\alpha\right)}{(2k-1)^2} = \frac{\pi}{4}\left(\frac{\pi}{2} - \alpha\right), \qquad \alpha \in [0, \pi]. \tag{14}$$

The above results can be combined to show that

$$\begin{aligned}
0 \;\leq\; & 4\pi^2 \sum_{k=1}^{\infty} \frac{\left|p_{2k-1} - q_{2k-1}\right|^2}{(2k-1)^2} \\
=\; & \sum_{k=1}^{\infty} \mathrm{E}\left(\frac{\cos\left((2k-1)\,\alpha(\Theta_P, \Theta_P^*)\right)}{(2k-1)^2} + \frac{\cos\left((2k-1)\,\alpha(\Theta_Q, \Theta_Q^*)\right)}{(2k-1)^2} \right. \\
& \qquad \left. - 2\,\frac{\cos\left((2k-1)\,\alpha(\Theta_P, \Theta_Q)\right)}{(2k-1)^2} \right) \\
=\; & \mathrm{E} \sum_{k=1}^{\infty} \frac{\cos\left((2k-1)\,\alpha(\Theta_P, \Theta_P^*)\right)}{(2k-1)^2} + \mathrm{E} \sum_{k=1}^{\infty} \frac{\cos\left((2k-1)\,\alpha(\Theta_Q, \Theta_Q^*)\right)}{(2k-1)^2} \\
& \qquad - 2\,\mathrm{E} \sum_{k=1}^{\infty} \frac{\cos\left((2k-1)\,\alpha(\Theta_P, \Theta_Q)\right)}{(2k-1)^2} \\
=\; & \frac{\pi}{4}\left(\mathrm{E}\left(\frac{\pi}{2} - \alpha(\Theta_P, \Theta_P^*)\right) + \mathrm{E}\left(\frac{\pi}{2} - \alpha(\Theta_Q, \Theta_Q^*)\right) - 2\,\mathrm{E}\left(\frac{\pi}{2} - \alpha(\Theta_P, \Theta_Q)\right) \right) \\
=\; & \frac{\pi}{4}\left(2\,\mathrm{E}\alpha(\Theta_P, \Theta_Q) - \mathrm{E}\alpha(\Theta_P, \Theta_P^*) - \mathrm{E}\alpha(\Theta_Q, \Theta_Q^*) \right),
\end{aligned}$$

which implies (12) and thereby (11). The first equality follows from (13), the second is justified by the dominant convergence theorem, the third follows from (14), and the final equality is immediate. The proof is complete.

References

Best, D. and Fisher, N. (1981). The bias of the maximum likelihood estimators of the von Mises-Fisher concentration parameters. *Communication in Statistics — Simulation and Computation*, **B10(5),** 493–502.

Bougeault, P., Binder, P., Buzzi, A., Dirks, R., Houze, R., Kuettner, J., Smith, R. B., Steinacker, R. and Volkert, H. (2001). The MAP special observing period. *Bulletin of the American Meteorological Society*, **82,** 433–462.

Candille, G. and Talagrand, O. (2005). Evaluation of probabilistic prediction systems for a scalar variable. *Quarterly Journal of the Royal Meteorological Society*, **131,** 2131–2150.

Case, J. L., Manobianco, J., Lane, J. E., Immer, C. D. and Merceret, F. J. (2004). An objective technique for verifying sea breezes in high-resolution numerical weather prediction models. *Weather and Forecasting*, **19,** 690–705.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, **74,** 829–836.

Done, J., Davis, C. A. and Weisman, M. (2004). The next generation of NWP: Explicit forecasts of convection using the Weather Research and Forecasting (WRF) model. *Atmospheric Science Letters*, **5,** 110–117.

Ebert, E. and McBride, J. L. (2000). Verification of precipitation in weather systems: Determination of systematic errors. *Journal of Hydrology*, **239,** 179–202.

Eckel, F. A. and Mass, C. F. (2005). Aspects of effective short-range ensemble forecasting. *Weather and Forecasting*, **20,** 328–350.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications, Vol. 2*, 2nd ed., John Wiley & Sons, New York, 669 pp.

Fisher, N. I. (1993). *Statistical Analysis of Circular Data*, Cambridge University Press, Cambridge, 277 pp.

Gneiting, T. and Raftery, A. E. (2004). Strictly proper scoring rules, prediction, and estimation. Technical Report no. 463, Department of Statistics, University of Washington, Seattle, Washington, USA.

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*, **33,** 1098–1118.

Gneiting, T., Larson, K., Westrick, K., Genton, M. G. and Aldrich, E. (2006). Calibrated probabilistic forecasting at the Stateline wind energy center: The regime-switching space-time (RST) method. *Journal of the American Statistical Association*, in press.

Gradshteyn, I. S. and Ryzhik, I. M. (1994). *Table of Integrals, Series, and Products*, 5th ed., Academic Press, New York, 1204 pp.

Grimit, E. P. (2001). Implementation and evaluation of a mesoscale short-range ensemble forecasting system over the Pacific Northwest. M.S. Thesis, Department of Atmospheric Sciences, University of Washington, Seattle, USA, 111 pp.

Grimit, E. P. (2004). Probabilistic mesoscale forecast error prediction using short-range ensembles. Ph.D. Dissertation, Department of Atmospheric Sciences, University of Washington, Seattle, USA, 146 pp.

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.

Hamill, T. M., Whitaker, J. S. and Mullen, S. L. (2005). Reforecasts: An important data set for improving weather predictions. *Bulletin of the American Meteorological Society*, in press.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, **15,** 559–570.

Jolliffe, I. T. and Stephenson, D. B. (2003). *Forecast Verification. A Practicioner's Guide in Atmospheric Science*, John Wiley & Sons, Chichester, 240 pp.

Mahrt, L. (1999). Stratified atmospheric boundary layers. *Boundary-Layer Meteorology*, **90,** 375–396.

Mass, C. F., Ovens, D., Westrick, K. and Colle, B. A. (2002). Does increasing horizontal resolution produce more skillful forecasts? *Bulletin of the American Meteorological Society*, **83,** 407–430.

Mass, C. F., Albright, M., Ovens, D., Steed, R., MacIver, M., Grimit, E., Eckel, T., Lamb, B., Vaughan, J., Westrick, K., Storck, P., Colman, B., Hill, C., Maykut, N., Gilroy, M., Ferguson, S. A., Yetter, J., Sierchio, J. M., Bowman, C., Stender, R., Wilson, R. and Brown, W. (2003). Regional environmental prediction over the Pacific Northwest. *Bulletin of the American Meteorological Society*, **84,** 1353–1366.

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

Nachamkin, J. E. (2004). Mesoscale verification using meteorological composites. *Monthly Weather Review*, **132,** 941–955.

Pocernich, M. (2006). R verification package, available at cran.r-project.org.

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.

Rife, D. L., Davis, C. A., Liu, Y. and Warner, T. T. (2004). Predictability of low-level winds by mesoscale meteorological models. *Monthly Weather Review*, **132,** 2553–2569.

Rife, D. L. and Davis, C. A. (2005). Verification of temporal variations in mesoscale numerical wind forecasts. *Monthly Weather Review*, **133,** 3368–3381.

Roulston, M. S. and Smith, L. A. (2002). Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, **130,** 1653–1660.

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

Stoelinga, M. T., Hobbs, P. V., Mass, C. F., Locatelli, J. D., Colle, B. A., Houze Jr., R. A., Rangno, A. L., Bond, N. A., Smull, B. F., Rasmussen, R. M., Thompson, G. and Colman, B. R. (2003). Improvement of microphysical parameterization through observational verification experiment. *Bulletin of the American Meterological Society*, **84,** 1807–1826.

Székely, G. J. (2003). $\mathcal{E}$-Statistics: The energy of statistical samples. Technical Report no. 2003–16, Department of Mathematics and Statistics, Bowling Green State University, Ohio, USA.

Toth, Z., Talagrand, O., Candille, G. and Zhu, Y. (2003). Probability and ensemble forecasts. Pp. 137–163 in *Forecast verification. A practicioner's guide in atmospheric science*, eds. I. Jolliffe and D. B. Stephenson. John Wiley & Sons, Ltd, Chichester, UK.

Unger, D. A. (1985). A method to estimate the continuous ranked probability score. Pp. 206–213 in Proceedings of the ninth conference on probability and statistics, Vir-

ginia Beach, Virginia. American Meteorological Society, Boston, USA.

Wang, X. and Bishop, C. H. (2005). Improvement of ensemble reliability with a new dressing kernel. *Quarterly Journal of the Royal Meteorological Society*, **131,** 965–986.

Wilks, D. S. (2006). *Statistical Methods in the Atmospheric Sciences*, 2nd ed., Elsevier Academic Press, Amsterdam, 627 pp.

Zhang, D.-L. and Zheng, W.-Z. (2004). Diurnal cycles of surface winds and temperatures as simulated by five boundary layer parameterizations. *Journal of Applied Meteorology*, **43,** 157–169.