

Hydrometeorological Short-Range Ensemble Forecasts in Complex Terrain. Part I: Meteorological Evaluation

DOUG MCCOLLOR

The University of British Columbia, and BC Hydro Corporation, Vancouver, British Columbia, Canada

ROLAND STULL

The University of British Columbia, Vancouver, British Columbia, Canada

(Manuscript received 2 August 2007, in final form 22 November 2007)

ABSTRACT

This paper addresses the question of whether it is better to include lower-resolution members of a nested suite of numerical precipitation forecasts to increase ensemble size, or to utilize high-resolution members only to maximize forecast details in regions of complex terrain. A short-range ensemble forecast (SREF) system is formed from three models running in nested configurations at 108-, 36-, 12-, and 4-km horizontal grid spacings. The forecasts are sampled at 27 precipitation-gauge locations, representing 15 pluvial watersheds in southwestern British Columbia, Canada. This is a region of complex topography characterized by high mountains, glaciers, fjords, and land-ocean boundaries. Matching forecast-observation pairs are analyzed for two consecutive wet seasons: October 2003–March 2004 and October 2004–March 2005. The northwest coast of North America is typically subject to intense landfalling Pacific cyclones and frontal systems during these months.

Using forecast analysis tools that are well designed for SREF systems, it is found that utilizing the full suite of ensemble members, including the lowest-resolution members, produced the highest quality probabilistic forecasts of precipitation. A companion paper assesses the economic value of SREF probabilistic forecasts for hydroelectric operations.

1. Introduction

The effect of horizontal resolution and ensemble size on a regional short-range ensemble forecast (SREF) system is assessed for probabilistic forecasts of 24-h accumulated precipitation in complex terrain. A multimodel SREF is used here rather than a multi-initial condition SREF. Model error, which is poorly understood and difficult to account for, has a larger impact on surface variables in the short range (Stensrud et al. 2000) than on free-atmosphere variables in the midrange.

Therefore, model error, being a large source of forecast uncertainty in short-range forecasts, must be accounted for to maximize SREF utility (Eckel and Mass 2005; Jones et al. 2007), particularly for mesoscale sensible weather elements.

In a recent SREF study of the surface elements mean sea level pressure (MSLP), temperature, and wind in the mountainous U.S. Pacific Northwest, Eckel and Mass (2005) found that a multimodel system outperformed a system that used variations of a single model. Eckel and Mass (2005) also found that an ensemble of unequally likely members can be skillful as long as each member occasionally performs well, and that the inclusion of finer-resolution models led to greater ensemble spread as smaller scales of atmospheric motion were modeled.

One of the keys to making skillful quantitative precipitation forecasts (QPF) in complex terrain is high-resolution modeling, to capture the orographic and variable-surface flux components of the precipitation distribution. Employing very high-resolution models is crucial to capturing variability at small scales (Hamill et al. 2000; Eckel and Mass 2005).

In Eckel and Mass (2005)'s study of an SREF system in the cool season over complex terrain, the impact of model error remained significant even where forecast

Corresponding author address: Doug McCollor, Dept. of Earth and Ocean Sciences, The University of British Columbia, 6339 Stores Rd., Vancouver, BC V6T 1Z4, Canada.
E-mail: doug.mccollor@bchydro.bc.ca

uncertainty was largely driven by synoptic-scale errors originating from analysis uncertainty. In complex terrain, many mesoscale weather phenomena, particularly cool season precipitation, are driven by the interaction between the synoptic-scale flow and underlying mesoscale topographic irregularities and boundaries.

A summary of results from Du et al. (1997), who analyzed a 25-member ensemble of QPFs, includes the finding that 90% of the improvement found in using an ensemble average was obtainable when using an ensemble size as small as 8–10 members. Further results from that paper report that an ensemble QPF from an 80-km grid model is more accurate than a single forecast from a 40-km grid model, and that SREF techniques can provide increased accuracy in QPFs even without further improvement in the forecast model system. Wandishin et al. (2001) found that even ensemble configurations with as few as five members can significantly outperform a higher-resolution deterministic forecast. An analysis of a global ensemble prediction system (EPS) for 24-h QPFs (Mullen and Buizza 2002) found that coarser-resolution, larger-member ensembles can outperform higher-resolution, smaller-member ensembles in terms of the ability to predict rare precipitation events.

This paper addresses the issue of SREFs for hydro-meteorological applications in complex terrain. The principal input into a hydrologic model, which produces forecasts of river stages or reservoir inflow, is the precipitation over the associated watershed (Krzysztofiwicz et al. 1993). The objective of the paper is to compare the skill of 24-h accumulated QPFs from various SREF configurations of three independent mesoscale NWP models run as multiple-resolution nested systems.

Section 2 describes the location, the models used in the SREF, and the dataset of verifying observations. Section 3 introduces the verification metrics for probabilistic forecasts, and section 4 gives the results. Conclusions are summarized in section 5.

2. Methodology

a. Location and observation data

Southwestern British Columbia is a region of complex topography characterized by high mountains, glaciers, fjords, and land–ocean boundaries. The region, lying between 48° and 51°N on the west coast of North America, is subject to landfalling Pacific cyclones and frontal systems, especially during the cool season months from October through March.

Precipitation data for this study were collected from a mesonetwork of 27 gauges within 15 small watersheds in southwestern British Columbia (see Fig. 1 for a ref-

erence map). Twenty-four of the stations are part of the hydrometric data collection program in support of reservoir operations for BC Hydro Corporation. The other three stations are operated by the Meteorological Service of Canada. Observations consist of the 24-h accumulation of precipitation of all types (where solid and mixed precipitation are measured in liquid form) based on the 1200 UTC synoptic hour.

Forecast system performance can be verified against model-based gridded analyses or against actual point observations. Gridded analyses offer more data for investigation through widespread coverage. However, gridded analyses are subject to model-based errors and smoothing errors, which are both of major concern in regions of complex terrain. Even though point observations are subject to instrumentation error and site-representativeness error, it is preferable to verify forecasts against observations as opposed to model analyses (Wilson 2000; Hou et al. 2001).

The observation stations used in this study range in elevation from 2 m AMSL at coastal locations to 1969 m AMSL in high mountainous watersheds. Because of the complex terrain in the study region and the fact that hydrologic models are often calibrated with station precipitation records, it was important in this study to verify forecasts against actual observations.

More information about the stations used in this study is shown in the summary hypsometric curve provided in Fig. 2 (watershed elevation area bands from all 15 watersheds are combined in this summary hypsometric curve). Two-thirds (67%) of the stations are below 500-m elevation (the MSL–500-m elevation band represents 13% of the area of the watersheds), while 18% of the stations are between 500- and 1500-m elevation, representing 44% of the area of the watersheds. The highest 15% of the stations, located between 1500 and 1969 m, represent 19% of the area of the watersheds.

Elevations above 2000 m, representing the highest 24% of the area of the watersheds, remain ungauged, with the highest peak in the watersheds at 2600 m.

Also shown in Fig. 2 is a cumulative total precipitation curve, as well as the average daily station precipitation from each station. This information reflects the nature of the nonuniform and rugged terrain, in that higher-elevation stations can be farther from the ocean and in the rainshadow of other mountains, and may lie within mountain ranges more parallel with the predominant flow, and therefore do not receive the highest precipitation amounts. **Lower-elevation stations, because of their open exposure to landfalling Pacific frontal systems and their orientation on mountain ranges perpendicular to the moist onshore flow, receive relatively high precipitation amounts.**

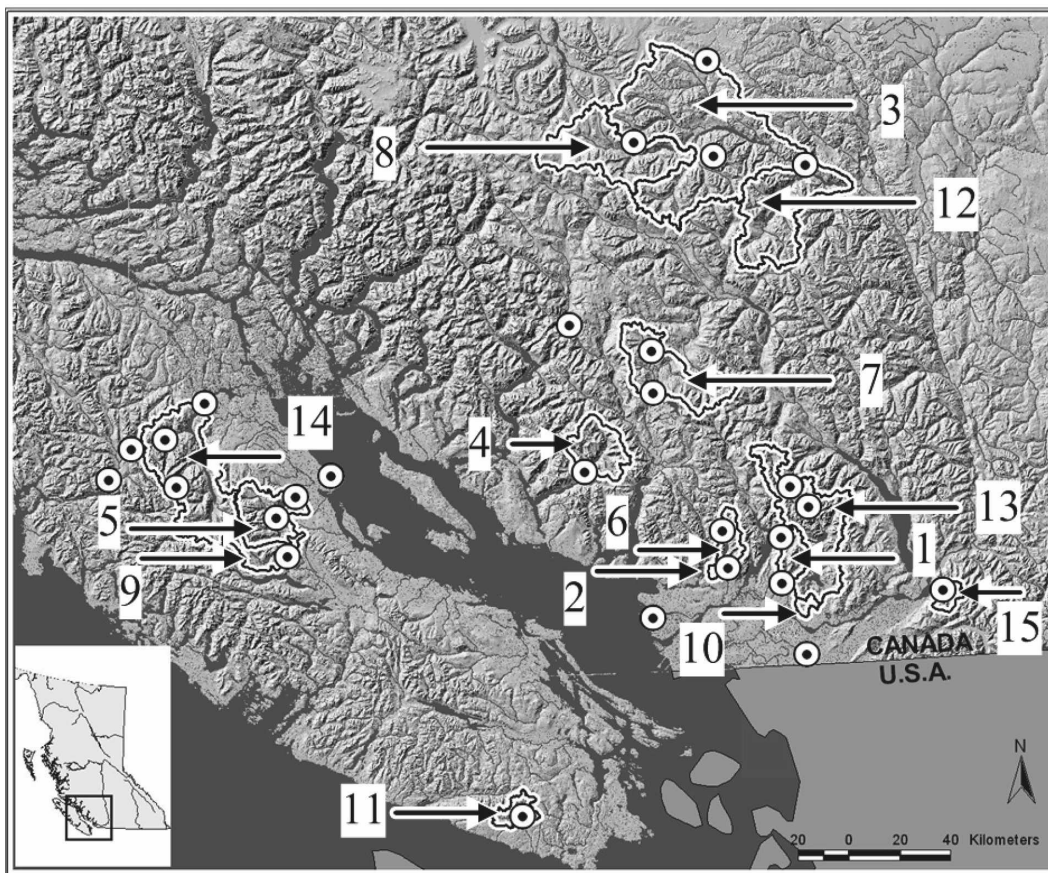


FIG. 1. Reference map for the area of study in southwestern BC, Canada. Twenty-seven weather station locations (dots) representing 15 watersheds (black lines) are depicted. Station elevation ranges from 2 to 1969 m above MSL. The reservoirs are numbered and listed by name in Table 2 of Part II.

b. Numerical models and study dates

The Geophysical Disaster Computational Fluid Dynamics Centre (GDCFDC), in the Department of Earth and Ocean Sciences at the University of British Columbia in Vancouver, runs a real-time suite of three independent nested limited-area high-resolution mesoscale models over the region of interest. For the study presented here, the coarse-resolution (108-km horizontal grid spacing) outer nests of these models were all initialized using the National Centers for Environmental Prediction (NCEP) North American Mesoscale (NAM) model (previously known as the Eta Model) at 90-km grid spacing. Time-varying lateral boundary conditions were also extracted from the NAM forecasts.

Ensemble forecasts of 24-h accumulated total precipitation were available from an archive of the three real-time limited-area model (LAM) runs over two consecutive wet seasons: October 2003–March 2004 and October 2004–March 2005. The Mesoscale Compressible Community model (MC2) is a fully compress-

ible, semi-implicit, semi-Lagrangian, nonhydrostatic mesoscale model (Benoit et al. 1997). One-way nesting is applied to produce model output at horizontal grid spacings of 108, 36, 12, 4, and 2 km. The fifth-generation Pennsylvania State University–National Center for Atmospheric Research Mesoscale Model (MM5) is a fully compressible, nonhydrostatic, sigma-coordinate model designed to simulate and predict mesoscale and regional-scale atmospheric circulations (Grell et al. 1994). The MM5 is run for the same five grids, but with two-way nesting. The Weather Research and Forecast (WRF) model (Dudhia et al. 1998; Skamarock et al. 2005) is a nonhydrostatic mesoscale model, run for three grids (108, 36, and 12 km) with two-way nesting applied.

The LAM gridded precipitation forecast values were interpolated to the precipitation gauge locations using a 2D cubic spline. The interpolation uses a 16-point (4×4) stencil, where the interpolated point is always inside the center square of the stencil.

It is important to incorporate uniquely different

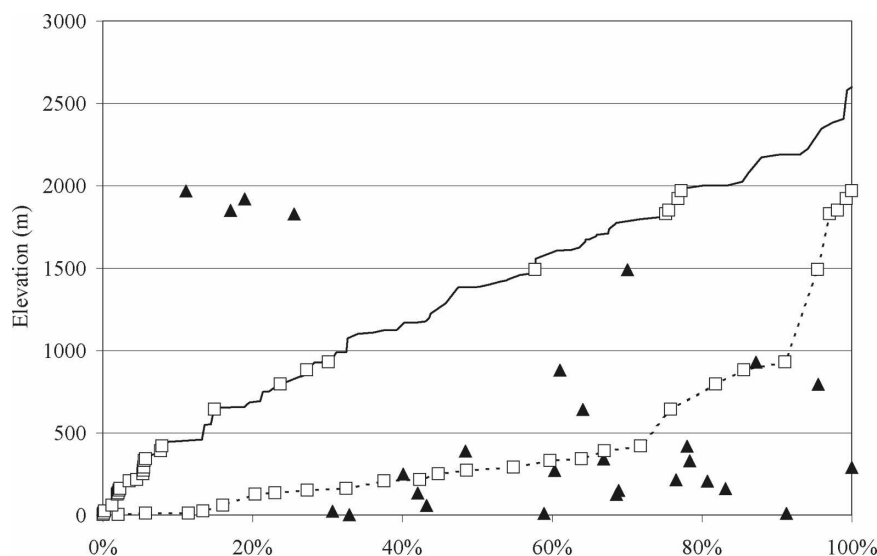


FIG. 2. Summary hypsometric curve (solid line) and cumulative total precipitation curve (dashed line). Individual stations (squares) are indicated on the curves. The total area of the 15 watersheds in this study is 7899 km². The total cumulative precipitation for all 27 stations over the study period is 87 324 mm. The triangles indicate the average daily station precipitation as a percentage of the station with the highest average daily precipitation (15 mm day⁻¹).

models in an SREF design, as opposed to a varied-model technique—a varied-model technique incorporates differing combinations of model physics and sub-grid-scale parameterization methods in a single model—to improve ensemble spread characteristics. Eckel and Mass (2005) found that even with extensive variations of a single model, a multimodel SREF design vastly outperformed a varied-model SREF design in representing model uncertainty.

Precipitation duration from individual extratropical weather systems in the cool season can vary from hours to days, so there is no obvious choice for choosing a particular time window for accumulation totals. The precipitation gauges available for this study record precipitation accumulation at 1-h intervals. However, the choice of verifying 24-h precipitation accumulation totals was made for two reasons. First, water managers often require daily time step inflow forecasts from hydrologic models. Second, lengthening the accumulation period beyond 1-, 3-, 6-, or 12-h accumulations monotonically and significantly increases the predictability (Wandishin et al. 2001; Stensrud and Yussouf 2007). For example, if a model forecast erred by delaying the start of a 12-h rainstorm by 4 h, then this rainfall would be missed entirely by some of the 1- and 3-h accumulation forecasts. The 6- and 12-h forecasts would capture part of the storm, while only the 24-h forecast would be considered entirely accurate. Of course, there will be instances in which the timing of storms will af-

fect even the 24-h accumulation window, but this impact is statistically lessened, and predictability is increased, by using the 24-h forecast over a large number of days.

The suite of nested LAMs is designed so that the lower-resolution models (108 km) encompass a large enough domain that weather features advected in from the boundaries are incorporated throughout the 60-h forecast window of the GDCFDC models. Successive higher-resolution models (36 and 12 km) incorporate smaller domains, eventually focusing on the region of interest with the highest-resolution models (4 and 2 km).

In building an SREF for hydrometeorological forecasting of precipitation, the question arises as to how best to incorporate these forecast models of multiple resolution into a viable EPS. A viable EPS requires enough ensemble members to reflect and characterize the spread of observations so that the probability distribution of the forecasts is, ideally, indistinguishable from the probability distribution of the observations.


Including lower-resolution models that incorporate a much larger domain may benefit the SREF design by increasing the spread in the ensemble. Nutter et al. (2004) showed that ensembles that use a larger domain produce greater spread than do smaller-domain ensembles because the use of periodically updated, coarse lateral boundary conditions can filter out short waves and reduce the amplitude of nonstationary waves en-

tering from the larger-domain model. Alternately, previous research (Mass et al. 2002) has indicated that forecast skill for precipitation in mountainous terrain improves as grid spacing decreases from 36 to 12 km. But Mass et al. (2002) found that verification scores generally degrade as resolution is further increased from 12 to 4 km as overprediction develops over windward slopes and crests of terrain (though for heavy precipitation amounts on windward slopes, the transition from 12 to 4 km enhances the forecast accuracy).

Lower-resolution models may not have the topographic resolution necessary to adequately resolve terrain-influenced precipitation patterns, but they have proven to improve model diversity and hence ensemble spread. Extremely high-resolution models may not incorporate accurate enough subgrid-scale parameterizations to characterize individual cloud precipitation processes. In addition, computer resource limitations prohibit running a large-member, high-resolution-only multimodel system.

Since nested-grid LAMs by design produce forecasts across a widely varying scale of resolutions, **this paper investigates the benefits of including multiple-resolution LAMs to increase ensemble size (hence improve spread) in an SREF system.** This approach of including multiresolution nested LAMs in an SREF is supported by Wandishin et al. (2001), who concluded that including less skillful members can add skill in a mixed-ensemble system.

Each of the three models was initialized at 0000 UTC and run for 60 h. The nesting nature of the models means that the 108-km run begins at the 0000 UTC initialization time. For the model with one-way nesting, the next 36-km nested run begins 3 h later, at 0300 UTC; the next 12-km nested run begins at 0600 UTC; and the 4-km nested run begins at 0900 UTC. Since all models are initialized with a dry start, there will be spinup errors in the early portion of each run, which lead to a dry bias in the model forecasts by not including precipitation that is occurring at initialization time.

To partially ameliorate the spinup problem, we ignore the first 12 h (0000–1200 UTC) and use forecast hours 12–36 as “day 1.” Note that the 2-km model forecast runs begin at 1200 UTC, so that forecasts from these very high resolution models would not incorporate any preforecast  period at all, leading to the maximum amount of dry bias originating from the dry start to the model. The forecasts from the period $T + 36$ to $T + 60$ of each model ensemble member were extracted to form the “day 2” forecast.

Not all forecast days were available due to model run failures; the 2-km runs of the MM5 and the MC2 were especially **susceptible to missing forecast days.** This,

plus the lack of any preforecast time incorporated into the 2-km model runs to alleviate spinup problems, contributed to the decision not to include 2-km forecast runs in the ensemble with the present operational configuration. Even without these 2-km forecasts, there remain 5027 forecast–observation pairs for the day-1 forecast and 4737 forecast–observation pairs for the day-2 forecast in this study.

Different configurations of an SREF system were constructed from the dataset to measure the influence of including lower-resolution and/or higher-resolution LAM ensemble members on the performance of the EPS. The full 11-member suite of ensemble members (all 11) included the MC2 (108, 36, 12, and 4 km), MM5 (108, 36, 12, and 4 km), and WRF (108, 36, and 12 km) models. A suite of eight ensemble members (hires8) including all but the lowest- (108 km) resolution models was constructed to measure the performance of the higher-resolution models only. Another suite of six medium-resolution ensembles (mres6) included the 36- and 12-km LAMs only. The mres6 configuration was included to examine the effect of excluding both the lowest- (108 km) and highest- (4 km) resolution models. The final suite (vhires5) included the five very highest-resolution LAMs only (12 and 4 km). Table 1 provides a summary of the SREF configurations.

3. Verification procedures

A single verification score is generally inadequate for evaluating all of the desired information about the performance of an SREF system (Murphy and Winkler 1987; Murphy 1991). Different measures, emphasizing different aspects and attributes of forecast performance, should be employed to assess the statistical reliability, resolution, and discrimination of an EPS. **A standardized set of evaluation methods, scores, and diagrams to interpret short- to medium-range ensemble forecasts is provided in Hamill et al. (2000)** and is incorporated here to assess an SREF system for hydro-meteorological forecasts. We use the degree of mass balance (DMB), mean error (ME), mean absolute error (MAE), mean square error (MSE), root-mean-square error (RMSE), Pearson correlation (r), linear error in probability space (LEPS), Brier skill score (BSS) including relative reliability and resolution terms, relative operating characteristic (ROC) curves derived from the hit rate (H) and from the false alarm rate (F), and rank histograms. A description of these evaluation methods is provided in the appendix. The reader is referred to Toth et al. (2003) or Wilks (1995) for comprehensive descriptions of the verification measures.

It is impossible to objectively assess the quality of an



TABLE 1. List of ensemble members for all four different SREF configurations. The “all11” configuration includes all three mesoscale model forecasts at all available horizontal resolutions. The “hires8” configuration includes all available members except the coarsest (108 km) members. The “mres6” configuration includes the medium-resolution mesoscale models at 12 and 36 km. The “vhires5” configuration consists of the five highest-resolution models only, at 4 and 12 km.

No.	Label	Description	All11	Hires8	Mres6	Vhires5
1	MM5–108	MM5 108-km resolution	✗			
2	MM5–36	MM5 36-km resolution	✗	✗	✗	
3	MM5–12	MM5 12-km resolution	✗	✗	✗	✗
4	MM5–4	MM5 4-km resolution	✗	✗		✗
5	MC2–108	MC2 108-km resolution	✗			
6	MC2–36	MC2 36-km resolution	✗	✗	✗	
7	MC2–12	MC2 12-km resolution	✗	✗	✗	✗
8	MC2–4	MC2 4-km resolution	✗	✗		✗
9	WRF–108	WRF 108-km resolution	✗			
10	WRF–36	WRF 36-km resolution	✗	✗	✗	
11	WRF–12	WRF 12-km resolution	✗	✗	✗	✗

individual ensemble forecast. Ensemble forecast systems must be verified over many cases. As a result, the scoring metrics described in the appendix are susceptible to several sources of noise (Hamill et al. 2000): improper estimates of probabilities arising from small-sized ensembles, insufficient variety and number of cases leading to statistical misrepresentation, and imperfect observations making true forecast evaluation impossible.

A complete evaluation of ensemble mesoscale forecast models involves many different verification methods and scores, observation criteria, interpretation of results, and definition of user needs. In fact, a succinct and satisfactory approach to ensemble mesoscale verification remains elusive (Davis and Carr 2000; Mass et al. 2002). Traditional objective approaches based on verification at fixed observing locations are greatly influenced by temporal and spatial errors, as well as deficiencies of the observing network. The value of high-resolution numerical forecasts is clearly user-dependant, and one verification system cannot address the needs of all forecast users. The verification in this study is designed for a hydrometeorological user needing to forecast water inflow into hydroelectric reservoirs.

4. Results

First, the performance of individual ensemble members is described and compared against other members. The reliability and resolution of an EPS’s performance are found next by using the Brier skill score. Probabilistic values are assigned to forecasts based on the number of ensembles exceeding a particular threshold, and are used to create ROC curves. Finally, we describe the extent to which individual ensemble members are

equally likely to verify by employing rank histogram diagrams.

a. Error performance of individual ensemble members

An important criterion in an EPS is that both forecasts and observations have similar distributions. Precipitation follows a distinctly asymmetrical distribution that is highly skewed to the right (tail toward high-precipitation values) with a physical limit of zero precipitation on the left. Such highly nonnormal distributions require analysis methods that are both resistant (not overly influenced by outliers) and robust (independent of the underlying distribution). In this study, we employ box-and-whisker diagrams and linear error in probability space (LEPS; see the appendix) to meet this requirement.

There are 11 available ensemble members, as defined in Table 1. Model and observation box-and-whisker percentiles are seen in Fig. 3 for day-1 forecasts and in Fig. 4 for day-2 forecasts. The median of all observations (raindays and nonraindays combined) is approximately 2 mm day^{−1} indicating that half of the station days in the cool season sample experienced precipitation of at least 2 mm day^{−1}. The box plots show percentiles only above the median because the level of detail in this portion of the distribution is of most hydrometeorological importance. The ordinate is logarithmic, to adequately display the full range of values.

The ensemble members show that the distributions for all models resemble the distribution for the observations at the 75th percentile, especially for the day-1 forecast. As the model distributions move into the climatologically rarer precipitation events (the 90th, 95th, and 99th percentiles, as well as the maximum event), a trend emerges showing that as the resolution increases,

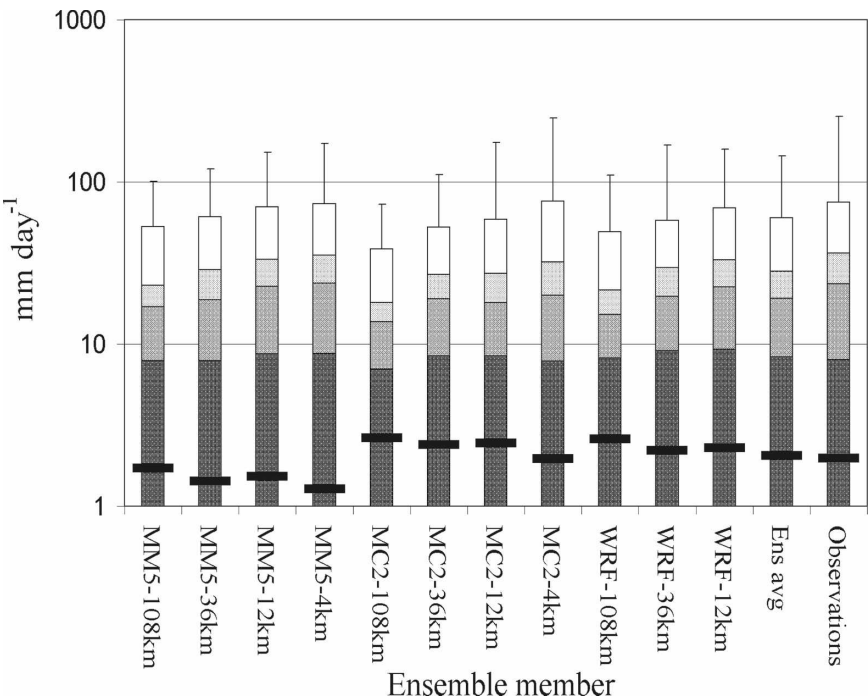


FIG. 3. Box-and-whisker plots for 11 members of the ensemble for the day-1 forecast of the 24-h precipitation amount. The ensemble average and associated observations are also included. The black horizontal bar indicates the median of each sample (raindays and nonrain-days combined). Increasing percentiles (lighter-colored bars) indicate the 75th, 90th, 95th, and 99th percentiles. The topmost “whisker” indicates the maximum value. The plots indicate that the distributions of the higher-resolution models more closely resemble the distribution of the observations for events greater than 2 mm day⁻¹.

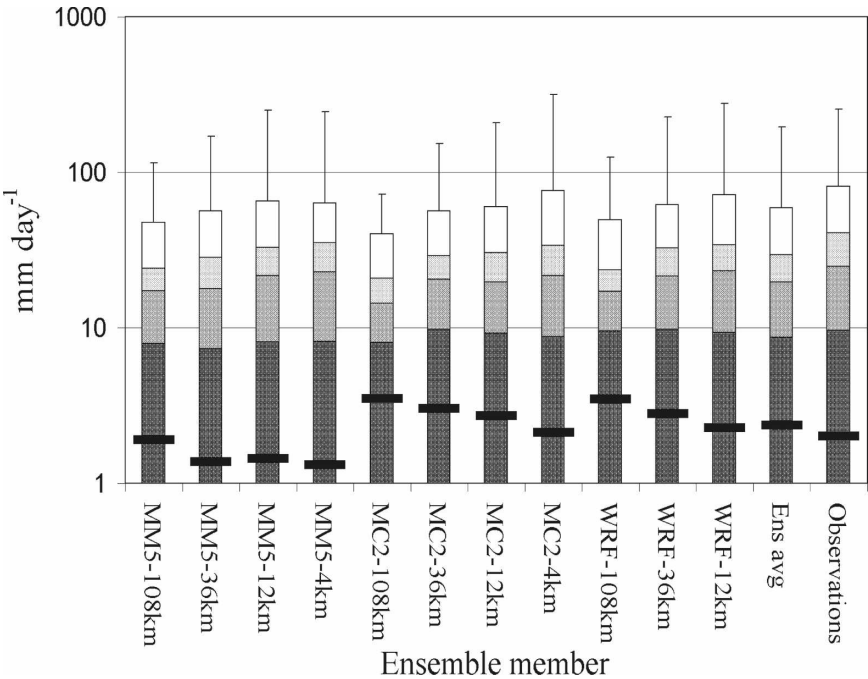


FIG. 4. Same as in Fig. 3, but for the day-2 forecast.

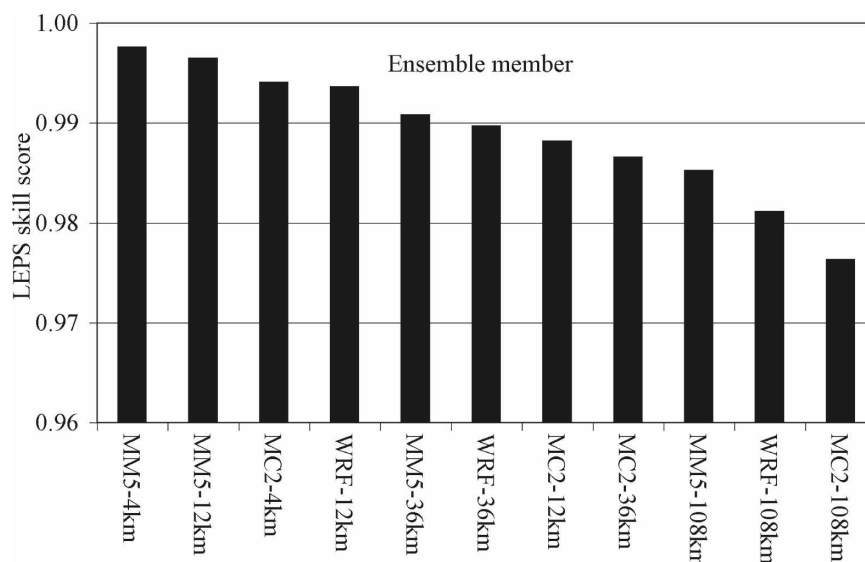


FIG. 5. The LEPS skill score for ensemble members for the day-1 forecasts. Values closer to 1 are better. The LEPS skill score compares the cumulative distributions of each model ensemble member with the corresponding distribution of the observations. Discrepancies in the distributions near the center of the distributions detract more from the LEPS skill score than do discrepancies in the extreme regions that correspond to rarer events. The coarsest-resolution members show a clear deterioration of skill compared to the finer-resolution members.

the model distributions more closely resemble the distribution of the observations. The two highest-resolution model distributions (4 km), MM5 and MC2 very closely resemble the distribution of the observations for the day-1 forecast. The three 12-km model distributions show reasonable agreement with the distribution of the observations, except for slight differences for the very rare events distinguished by the 99th percentile and maximum value categories. The lower-resolution model distributions, at 108 and 36 km, do not capture the rare events (90th percentile and higher) as well as the high-resolution model distributions for either the day-1 or day-2 forecasts. For the day-2 forecast, the lowest-resolution model distributions diverge somewhat from the distribution of the observations even at the 75th percentile.

A LEPS skill score with the climatological median as a reference is shown in Fig. 5 for the day-1 forecasts and in Fig. 6 for the day-2 forecasts. In terms of LEPS, the 108-km models perform the poorest, followed in general by the 36-km models for both forecast days. The 12- and 4-km models exhibit the best results.

In addition to LEPS, standard summary measures of examining the error performance of individual ensemble members in this study have been calculated, and include the degree of mass balance (DMB), Pearson product moment correlation, mean error, mean absolute error, and root-mean-square error (see the appen-

dix for a summary of the equations used for the meteorological statistical analysis of the individual ensemble members).

DMB values indicate that all 108- and 36-km resolution models exhibit an underforecast tendency (DMB < 1) averaged over all stations in this region of complex terrain, for both forecast days 1 and 2 (see Fig. 7). DMB values exhibit a clear trend of improving the mass balance for the higher-resolution models, so that the MM5-12 km, WRF-12 km, MM5-4 km, and MC2-4 km all exhibit nearly perfect mass balance in the day-1 forecasts. Day-2 forecasts follow a trend similar to their day-1 counterparts in terms of improving mass balance with the higher-resolution models; however, even the highest-resolution models show a slight tendency to underforecast in the day-2 time frame.

Mean error improves dramatically with higher resolution. The finest-resolution models exhibit negligible mean error (less than 0.5 mm day^{-1}) for the day-1 forecasts. The mean error shows the same trend toward underforecasting for the coarser-resolution models as exhibited by DMB. Day-1 forecasts show improvement in mean error over the day-2 forecasts.

The mean absolute error of the forecasts is highly consistent across models and resolutions, between 5 and 7.5 mm day^{-1} . Day-2 forecasts show consistently higher mean absolute error than corresponding day-1 forecasts.

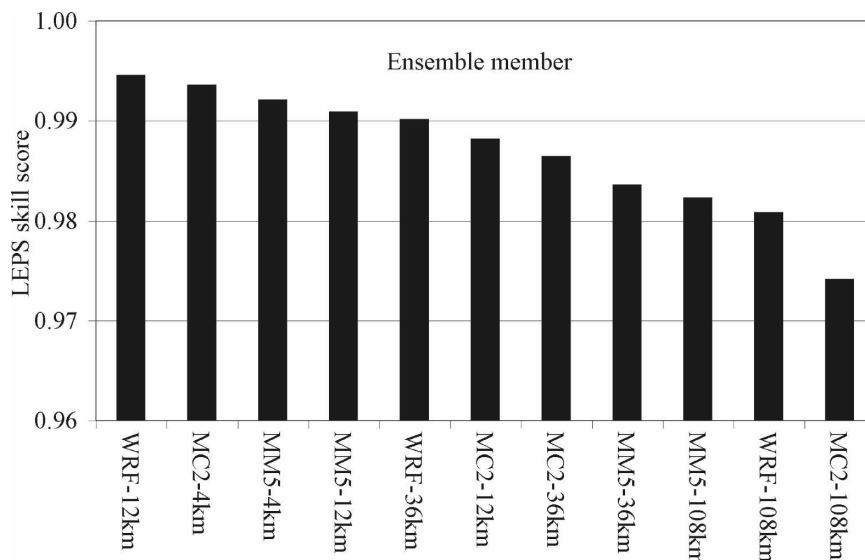


FIG. 6. The LEPS skill score for ensemble members for the day-2 forecasts. Values closer to 1 are better. The coarsest-resolution members show a clear deterioration of skill compared to the finer-resolution members, as for the day-1 forecasts. In general, the day-2 forecasts are slightly less skillful than are the day-1 forecasts in terms of the LEPS skill score.

The mean error and mean absolute error results can be explained in light of the definition of mean error as an overall measure of systematic bias in the forecasts (Jolliffe and Stephenson 2003). NWP model forecasts contain systematic biases due to imperfect model physics, initial conditions, and boundary conditions (Cheng and Steenburgh 2007). Also, NWP models prognose fields at gridpoint locations that represent a volume average, yet observations are taken at a specific location. Cheng and Steenburgh (2007) also make the point that bias results from differences between model and actual station elevations, plus nearby topographic influences too small to be represented in the model. These systematic biases are model-resolution dependent as different-resolution versions of a model will represent local topography differently. Higher-resolution models should represent local topography better and have less bias than lower-resolution models. This is evident in the results shown in Fig. 7. Systematic bias can be addressed and corrected for via postprocessing of the forecasts (Cheng and Steenburgh 2007; Yussouf and Stensrud 2007; McCollor and Stull 2008a).

Random error has positive variance but zero mean. Therefore, random error is a component of mean absolute error, but not the mean error. Random error is a function of the model physics and parameterization schemes and is driven by how well the forecast model performs in representing the atmosphere. Random error may be model dependant but is not resolution dependant. Random error cannot be improved by post-

processing model forecasts. The core model must be improved to diminish random error. Therefore, no significant variations in mean absolute error are evident among the different resolutions in our results. Random error can be reduced by making ensemble averages.

The root-mean-square error of the forecasts is highly consistent across all models and resolutions, between 10 and 15 mm day⁻¹. Day-2 forecasts exhibit consistently higher rms error than do the corresponding day-1 forecasts.

Correlation between forecasts and observations is highly consistent among models, resolutions, and forecast days. The Pearson product moment correlation is between 0.65 and 0.75 for all models for both day-1 and day-2 forecasts. The finer-resolution models (12 and 4 km) tend to show a higher correlation between forecasts and observations than do the coarser-resolution models (108 and 36 km). Day-1 forecasts exhibit slightly higher correlation than the corresponding day-2 forecasts.

In summary, an intercomparison among all models and resolutions indicates a definite trend toward higher similarity among forecast–observation distributions for the finer-resolution models. A trend toward better mass balance and improving mean error statistics is found in the finer-resolution models. Mean absolute error, root-mean-square error, and correlation show little variation among different-resolution models, but do show improving error statistics moving from the day-2 forecast to the day-1 forecast.

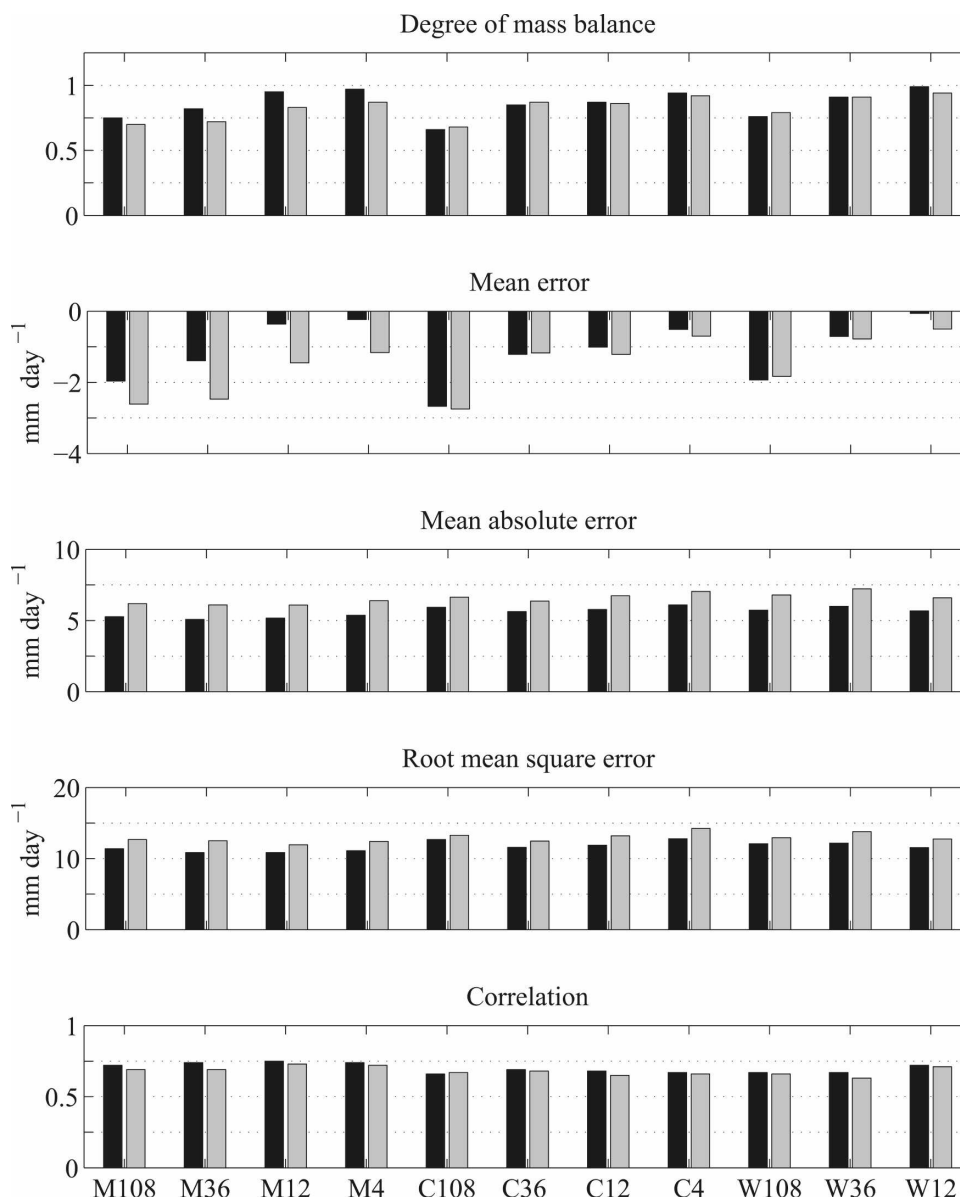


FIG. 7. Error characteristics of individual ensemble members for the full 11-member ensemble for day-1 (black) and day-2 forecasts (gray). Here, M refers to the MM5 model, C the MC2 model, and W the WRF model. The associated number is the model resolution (km).

The differences between the coarser-resolution and finer-resolution models vary across different metrics, indicating that each of the 11 members has the potential to contribute to improving the quality of an ensemble prediction system for 24-h precipitation. As described in section 2, the full 11-member ensemble was subdivided into smaller ensembles to test the contribution of the coarse-resolution and fine-resolution members relative to the full suite of ensemble members. The following sections employ verification measures designed for an EPS to examine the contribution of the coarse- and

fine-resolution members in building an operational SREF system.

b. Brier skill score: Resolution and reliability

Brier skill score results (BSS_{∞} ; see the appendix), incorporating an adjustment factor so that the different SREF systems composed of different size ensembles can be effectively compared, and including relative reliability and relative resolution components, are shown in Fig. 8 (5 mm day⁻¹ precipitation threshold), Fig. 9 (10 mm day⁻¹ threshold), Fig. 10 (25 mm day⁻¹ thresh-

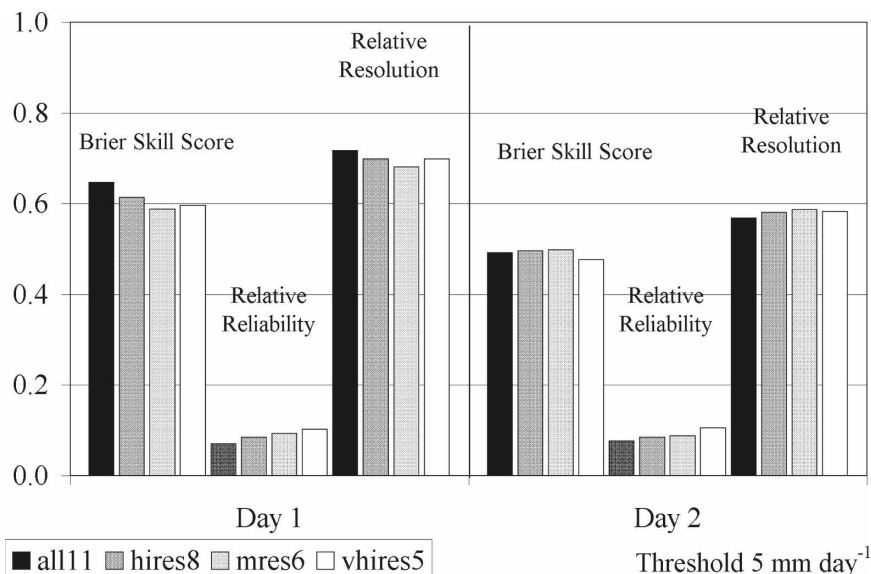


FIG. 8. BSS (1 is perfect), relative reliability (0 is perfect), and relative resolution (1 is perfect) for the SREFs for forecast days 1 and 2. The precipitation threshold is 5 mm day⁻¹. The configuration encompassing all 11 ensemble members (from 108 km down to 4 km for all three models) performs better for the day-1 forecast than do the other configurations with fewer ensemble members. The four SREF configurations show similar Brier skill scores for the day-2 forecast. Note that resolution deteriorates in the day-2 time frame while reliability is consistent for both the day-1 and day-2 forecasts.

old), and Fig. 11 (50 mm day⁻¹ threshold). The reference system for the Brier skill score is the climatological forecast in which the probability of the event is derived from the average of all observations in the sample.

The full 11-member ensemble shows the best Brier

skill score (BSS_{∞} , consisting of the best reliability and resolution components) for all four thresholds for the day-1 forecast and for the higher precipitation thresholds for the day-2 forecast. For the 5 mm day⁻¹ threshold day-2 forecast, the 11-member ensemble shows skill that is similar to that of the fine-resolution eight-

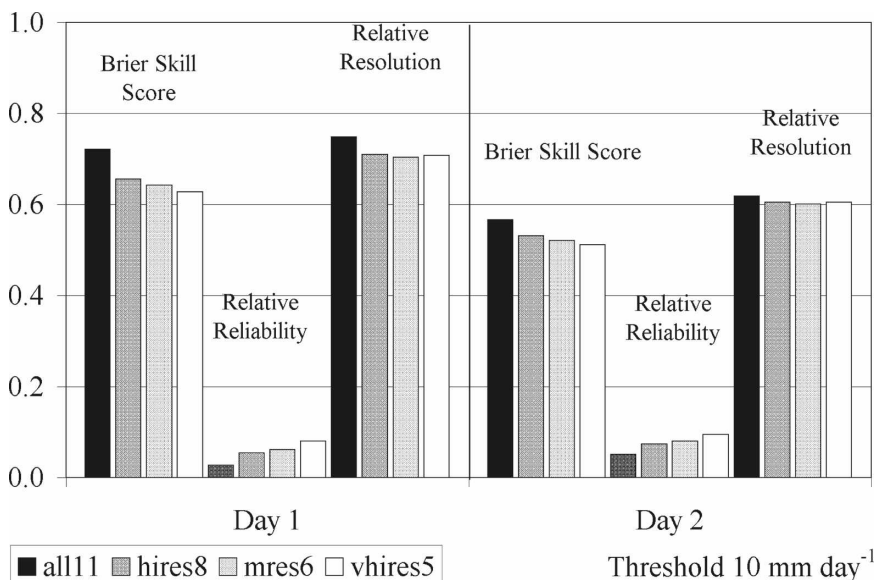


FIG. 9. Same as in Fig. 8, but for the 10 mm day⁻¹ precipitation threshold. The configuration encompassing all 11 ensemble members performs best for both the day-1 and day-2 forecasts at this threshold.

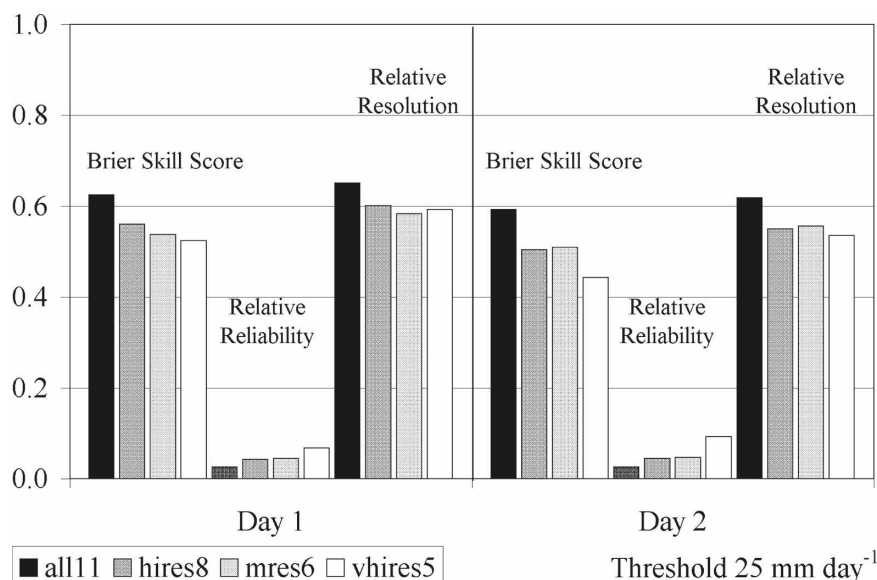


FIG. 10. Same as in Fig. 8, but for the 25 mm day⁻¹ precipitation threshold. The configuration encompassing all 11 ensemble members performs best for both the day-1 and day-2 forecasts at this threshold.

member and medium-resolution six-member SREF systems.

The very fine-resolution five-member SREF performs the worst in terms of Brier skill score, for all thresholds and for both forecast days, while the fine-resolution eight-member SREF is generally second best and the medium-resolution 6-member SREF is generally third best. Thus, increasing membership size, even at the expense of including coarser-resolution mem-

bers, is advantageous in terms of reliability and resolution across a wide range of 24-h precipitation thresholds, for this case study. All SREF configurations show skill relative to climatology.

The reliability diagrams for the full 11-member ensemble for different 24-h precipitation thresholds are shown in Fig. 12 for the day-1 forecasts, and in Fig. 13 for the day-2 forecasts. The SREF exhibits reasonably good reliability as indicated by the reliability diagrams.

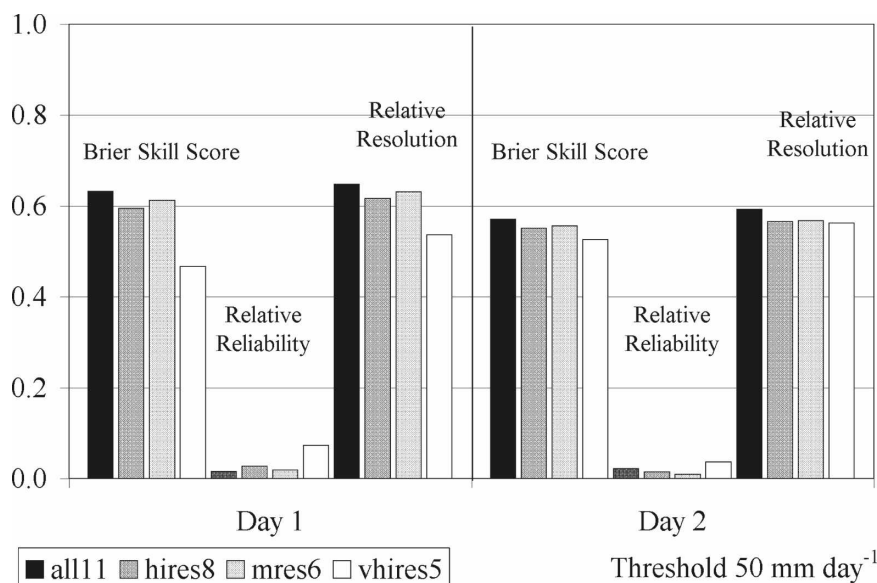


FIG. 11. Same as in Fig. 8, but for the 50 mm day⁻¹ precipitation threshold. The configuration encompassing all 11 ensemble members performs best at this threshold.

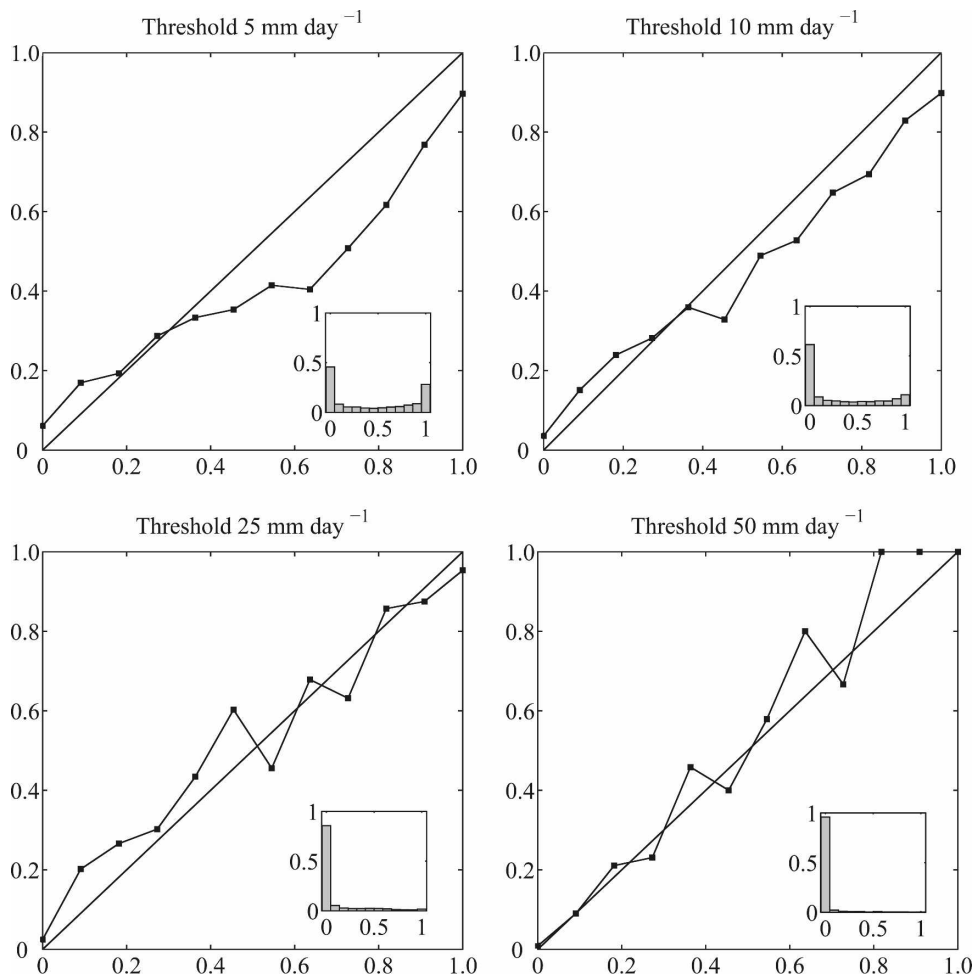


FIG. 12. Reliability diagrams from the full 11-member ensemble for the day-1 forecasts for precipitation thresholds of 5, 10, 25, and 50 mm day⁻¹. The abscissas in these graphs show the forecast probability, and the ordinates are the observed relative frequency. The straight diagonal line indicates perfect reliability. The inset diagrams depict sharpness, where the abscissa is forecast probability and the ordinate is the relative frequency of use.

The reader is referred to Hamill (1997) or Wilks (1995) for a display of a variety of hypothetical reliability diagrams useful in gauging different degrees of reliability and sharpness. The forecasted probabilities match the observed relative frequencies fairly well, though the 5 and 10 mm day⁻¹ thresholds show slight overforecasting at the higher probabilities for both forecast days 1 and 2. The zigzag pattern of the distribution of points in the rarer 25 and 50 mm day⁻¹ event threshold reliability diagrams indicates that these results exhibit signs of small sample size. Reliability diagrams for higher thresholds, 75 and 100 mm day⁻¹ (not shown), exhibit a highly erratic pattern due to the small statistical sample realized at these high precipitation thresholds.

Reliability diagrams for the other SREF systems are not shown, but the relative reliability for the full 11-

member system (Figs. 8–11) tends to be better than that for the other lower-member systems.

The SREF system exhibits a high degree of sharpness, especially at the higher precipitation thresholds of 10, 25, and 50 mm day⁻¹ (see insert histograms in Figs. 12 and 13). The high degree of sharpness indicated by the forecast systems ensures that the forecasts do not cluster near the climatological mean.

c. ROC curves

The transformed ROC curves for the four SREF systems for different 24-h precipitation thresholds are shown in Figs. 14 and 15. ROC area skill scores greater than 0.4 indicate reasonable and useful discriminating ability; values 0.6 or higher indicate good discriminating ability; and ROC area skill scores of 0.8 indicate excel-

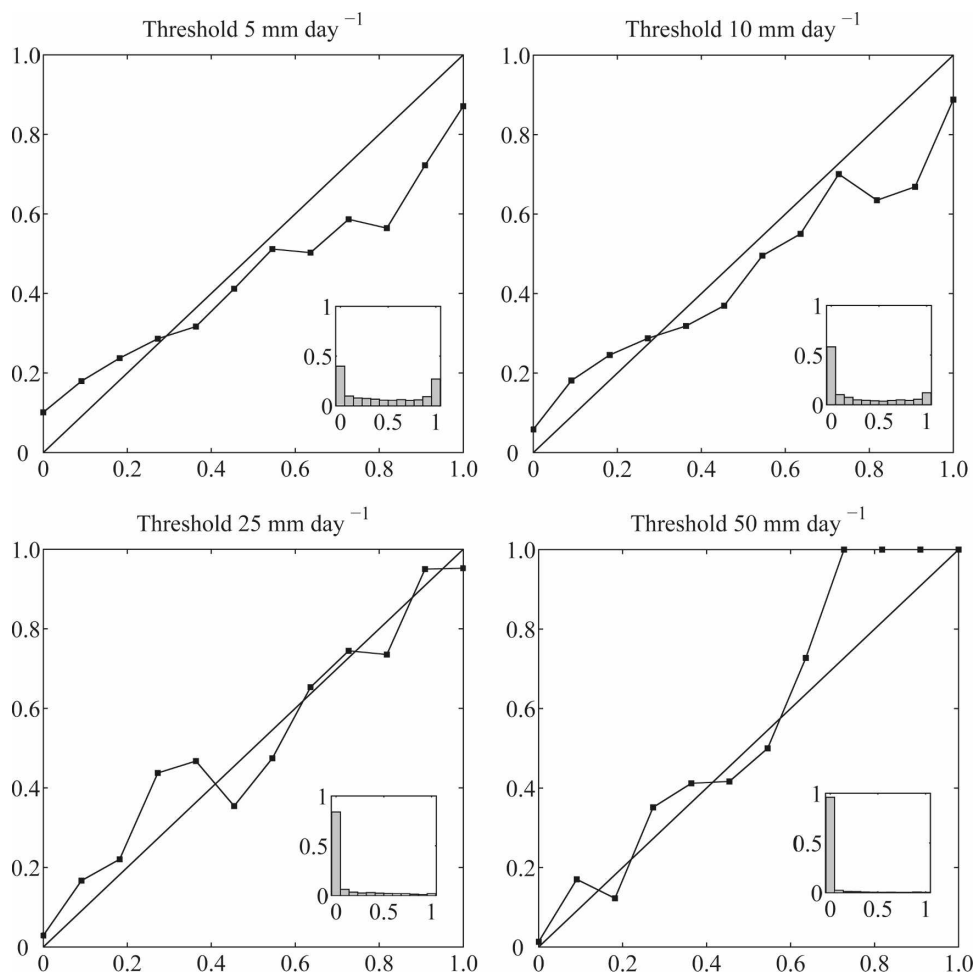


FIG. 13. Same as in Fig. 12, but for the day-2 forecasts.

lent discriminating ability (Buizza et al. 1999; Bright et al. 2005; Yuan et al. 2005; Stensrud and Yussouf 2007).

Table 2 examines the ability of the EPS to discriminate among events by comparing ROC area skill scores for all four different SREF configurations in detail. All four SREF configurations are able to discriminate between events by exhibiting ROC skill scores greater than 0.6 for each threshold, for both forecast days. ROC skill scores increase as the precipitation threshold increases for the day-1 forecasts, with little difference among the different SREF configurations. ROC skill scores are lower for the day-2 forecasts than for the day-1 forecasts, indicating a deterioration in discriminating ability with forecast period. Discriminating ability, as measured by ROC skill scores for the day-2 forecast, also increases with precipitation threshold up to the 25 mm day⁻¹ threshold. The forecasts for the 50 mm day⁻¹ threshold for the day-2 forecast indicate a marked decrease in discriminating ability compared to the 25 mm day⁻¹ threshold, the opposite trend to the

day-1 forecasts (also evident from the ROC curves for the day-2 forecasts in Figs. 14 and 15).

The plotted points of the actual ROC curves tend to cluster more toward the bottom left-hand corner of the curve for higher precipitation thresholds; therefore, the full transformed curves (see the appendix) contain more area and reflect higher skill scores. This does not hold true for the 50 mm day⁻¹ threshold forecasts on day 2 as the SREF systems lose discriminating ability at these high-threshold, rare events in the longer-range day-2 time frame. These results are consistent with an analysis of transformed ROC curves performed by Wilsson (2000).

d. Equal likelihood

The rank-histogram (Talagrand) diagrams for the day-1 and day-2 forecasts for the full 11-member ensemble are shown in Fig. 16. The measure of the flatness of the histograms, δ , for the different SREF configurations is given in Table 3.

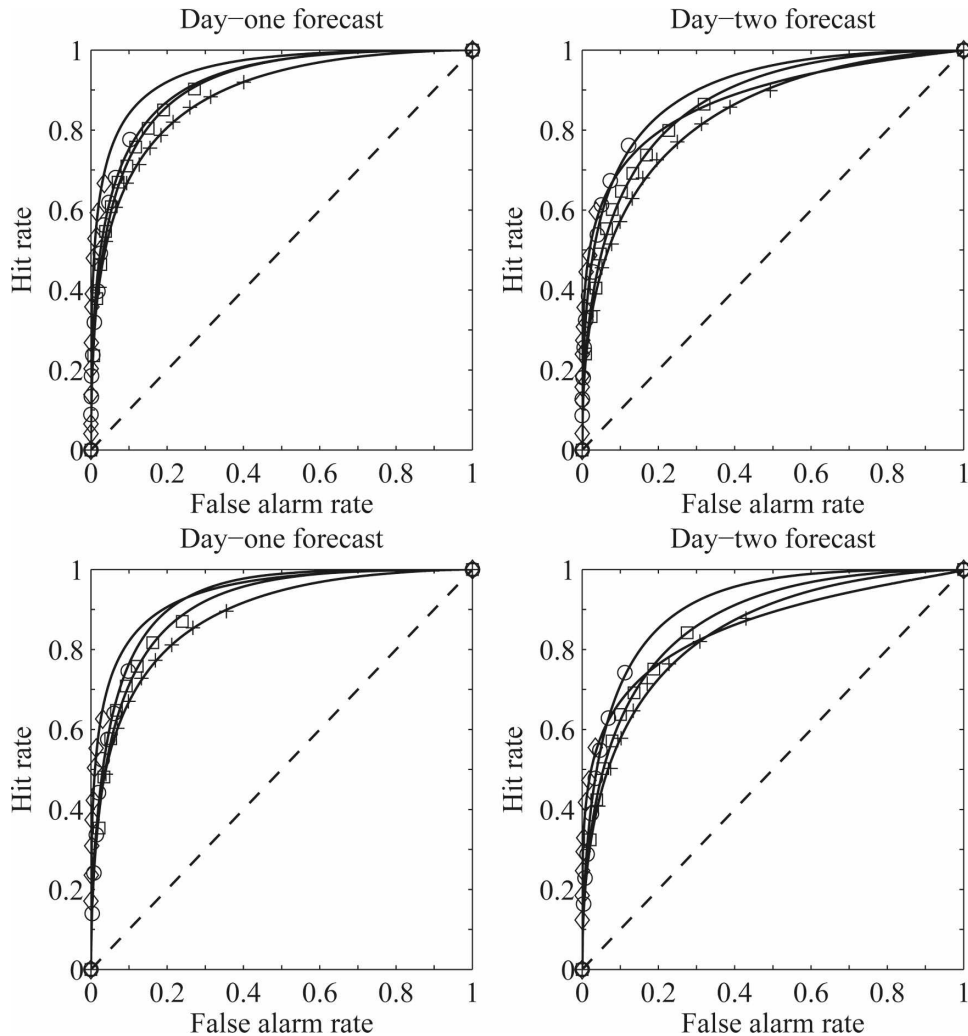


FIG. 14. ROC curves for (top) the full 11-member and (bottom) high-resolution 8-member ensemble prediction systems for precipitation thresholds at 5 (crosses), 10 (squares), 25 (circles), and 50 mm day⁻¹ (diamonds): (left) day-1 and (right) day-2 forecasts. The area under the ROC curves provides a measure of discrimination skill. Table 2 provides a summary of the ability of the SREF system to discriminate among events for all four different SREF configurations.

The rank histogram for the full 11-member ensemble exhibits a U shape, indicating that the spread of the ensemble is underdispersive for 24-h precipitation forecasts. Eliminating ensemble members, as is done in the other three SREF configurations, can only exacerbate the EPS problem of underdispersal. Table 3 shows that $\delta \gg 1$ for all SREF configurations for both forecast days 1 and 2, indicating a lack of spread among all ensembles. The normalized reliability index (RI), shown in Table 4, also shows that the full 11-member ensemble provides the best reliability and that reliability does, in fact, improve with increasing ensemble size.

Poor ensemble spread (underdispersion) is a well-known limitation of an EPS, especially SREFs that

forecast precipitation and other surface weather elements. However, ensembles still possess skill equal to or better than a single deterministic forecast run at higher resolution (Wandishin et al. 2001). Eckel and Mass (2005) show, through a display of rank histograms, that model diversity plays a much greater role for surface sensible weather elements (wind speed and temperature) than for other synoptic variables (50-kPa height and MSLP).

5. Conclusions

Forecasts of 24-h precipitation generated by a short-range ensemble forecast system have been analyzed to

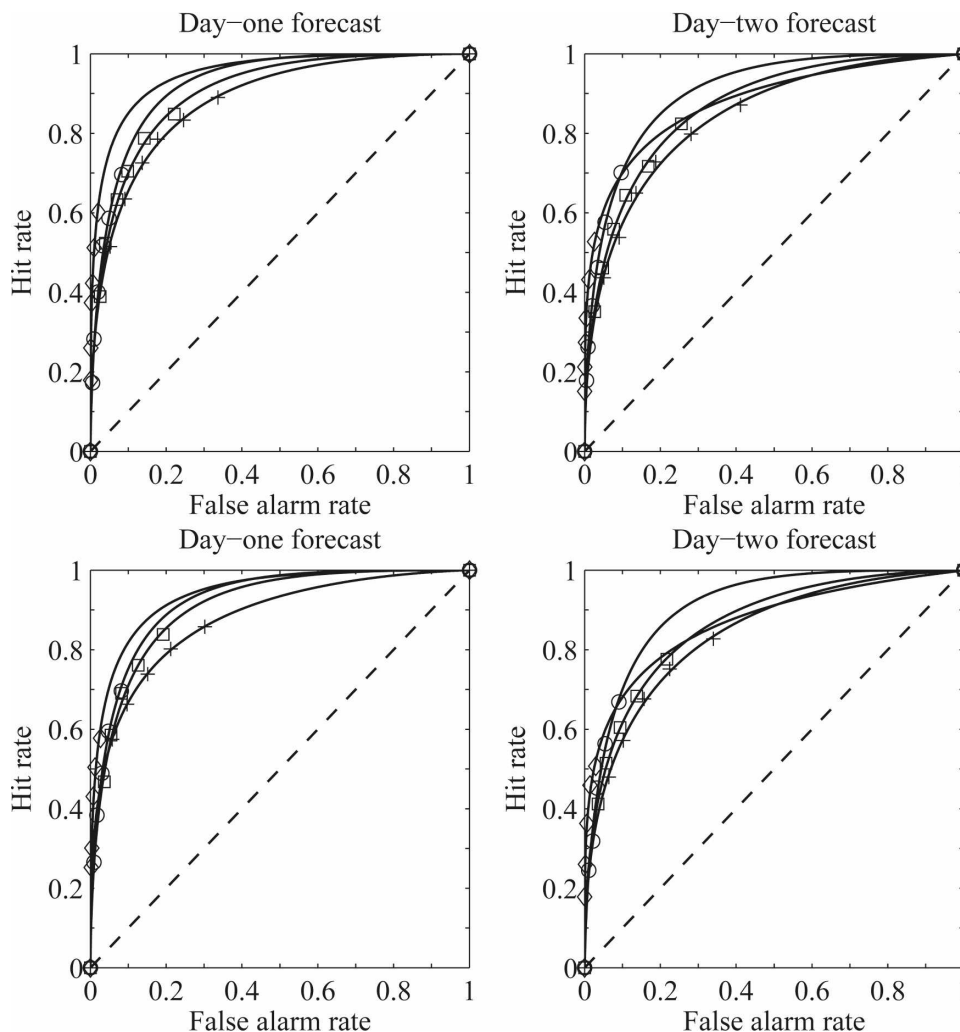


FIG. 15. Same as in Fig. 14, but for the (top) medium-resolution six-member and (bottom) very high-resolution five-member ensemble prediction systems.

study the usefulness of such forecasts in a hydrologic ensemble prediction system. Precipitation is the key forecast parameter in hydrologic modeling for the 15 small pluvial watersheds included in this study.

In this study, three different limited-area mesoscale models (MC2, MM5, and WRF) are run in a nested fashion, generating forecasts at up to four different telescoping resolutions (horizontal grid spacings of 108, 36, 12, and 4 km). This paper addresses the question of whether it is beneficial to include the low- (108 km) and high- (4 km) resolution members to increase the size of the ensemble, knowing the limitations of mesoscale models at very low (poor topographic representation) and very high (inadequate subgrid-scale parameterization) resolutions.

The results of this study for precipitation in complex terrain show that it is best to include all available reso-

lution models in the SREF configuration, even at the expense of including coarser-resolution members that exhibit higher errors and lower correlation with observations. The finer-resolution members exhibited the lowest error characteristics of the suite of individual members, and should definitely be included in the ensemble. The benefit of including more members includes improved reliability and resolution components of the Brier skill score. Including more coarse-resolution members does not inhibit discrimination, as determined by ROC scores. In terms of equal likelihood, the full 11-member ensemble exhibits underdispersion of the ensemble (a trait common among other ensemble prediction systems), as did all SREF configurations examined in this paper. However, the full 11-member ensemble did exhibit the greatest reliability relative to the other tested configurations.

TABLE 2. Values of ROC area skill scores for all four different SREF configurations for forecast days 1 and 2. Values of ROC area skill scores closer to 1 are better. ROC area skill scores greater than 0.6 indicate good discriminating ability; scores 0.8 or higher indicate excellent discriminating ability. All four SREF configurations exhibit good to excellent discriminating ability. Discrimination generally increases with increasing precipitation threshold, except for the 50 mm day⁻¹ threshold on day 2. Discrimination deteriorates moving from the day-1 to day-2 forecasts.

Day-1 forecast				
SREF configuration	5 mm	10 mm	25 mm	50 mm
All11	0.77	0.82	0.84	0.89
Hires8	0.77	0.82	0.85	0.88
Mres6	0.76	0.81	0.85	0.90
Vhires5	0.75	0.82	0.84	0.88

Day-2 forecast				
SREF configuration	5 mm	10 mm	25 mm	50 mm
All11	0.68	0.74	0.80	0.75
Hires8	0.69	0.74	0.82	0.71
Mres6	0.69	0.74	0.81	0.76
Vhires5	0.68	0.73	0.82	0.73

Day-2 forecasts exhibit a definite deterioration in error characteristics, Brier skill score, and discrimination over day-1 forecasts. However, the variations are generally slight and day-2 forecasts remain skillful in terms of the stated probabilistic verification measures.

The evidence gleaned from previous SREF studies and the results presented here indicate that SREF performance, in terms of skill and dispersion, may be increased by incorporating certain model run procedures. The spinup problem of mesoscale models can be avoided or lessened by increasing the data assimilation-preforecast period for the higher-resolution model runs. Including the very fine (2 km or finer) resolution nested model runs would generate increased ensemble dispersion at the smaller scales, but the error characteristics of these extremely fine-resolution models need to be examined.

Precipitation forecasting on the scale of small pluvial watersheds can be a daunting task, especially on the west coast of North America where the upwind Pacific data void introduces greater uncertainty (Hacker et al. 2003; McMurdie and Mass 2004; Spagnol et al. 2004) into the initial conditions of numerical weather forecasts than for regions in the center and on the eastern side of the continent. Forecast precipitation amounts over small watersheds depend largely on the storm track, and a shift of a few tens of kilometers to the north or south can make a significant difference between a “correct forecast” and a “miss” or “false alarm” of high precipitation that leads to high inflow (Weber et al. 2006).

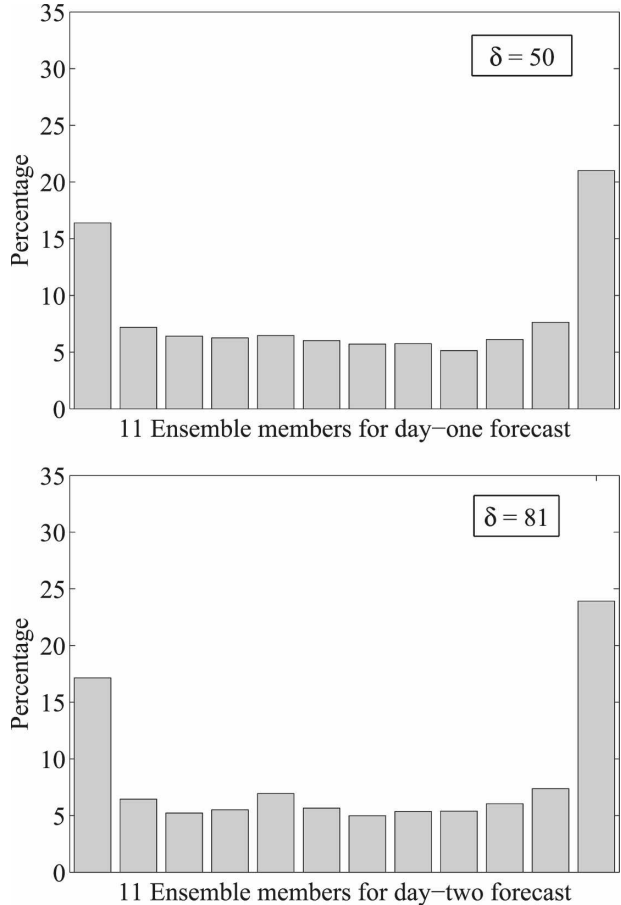


FIG. 16. Rank histogram diagrams for (top) day-1 and (bottom) day-2 forecasts for the full 11-member SREF configuration. The values of $\delta \gg 1$ for each forecast day indicate that the histogram is not flat.

Support for the development of ensemble hydrometeorological prediction systems is a predecessor to the goal of a skillful coupled precipitation-runoff ensemble forecast system. This paper is Part I of a two-part evaluation study that recent investigations (Yuan et al. 2005) confirm must be performed for atmospheric variables that have historically not been scrutinized. McCollor

TABLE 3. Values of flatness, δ , for all four SREF configurations for forecast days 1 and 2, where $\delta = 1$ indicates a perfectly flat rank histogram. Values of $\delta \gg 1$ indicate underdispersion in the ensemble forecasts. All four EPS configurations exhibit underdispersion, a common trait among mesoscale SREFs.

SREF configuration	Day-1 forecast	Day-2 forecast
All11	50	81
Hires8	51	83
Mres6	40	66
Vhires5	42	76

TABLE 4. Values of the reliability index (RI) for all four SREF configurations for forecast days 1 and 2. Lower values of RI are better, with a value of zero indicating perfect flatness of the rank histogram. Here, RI is normalized so that ensembles of different size can be compared with one another. The full 11-member ensemble shows the best RI values. Reliability, as measured by this reliability index, improves with increasing ensemble size and shorter forecast period.

SREF configuration	Day-1 forecast	Day-2 forecast
All11	8.54	10.5
Hires8	10.7	12.9
Mres6	11.7	13.8
Vhires5	13.9	16.7

and Stull (2008b, hereafter Part II) concludes this series of papers by incorporating user-dependant economic analyses to ensure a full suite of forecast evaluation methodologies are addressed in determining the usefulness of the SREF system.

Acknowledgments. The authors thank BC Hydro Corporation and the Meteorological Service of Canada for providing the precipitation observations necessary to perform this study. In addition, the Geophysical Disaster Computational Fluid Dynamics Centre (GDCFDC), located within the Department of Earth and Ocean Sciences at the University of British Columbia (UBC), is acknowledged for providing the numerical model forecasts used in this study. In particular, we thank Mr. George Hicks II for his efforts in making the numerical forecast information available to us. The GDCFDC gratefully acknowledges the following organizations for computer-purchase grants: Canadian Foundation for Innovation, British Columbia Knowledge Development Fund, and UBC. Additional grant support was provided by the Canadian Natural Sciences and Engineering Research Council, and Environment Canada. At UBC the WRF model was run operationally by Henryk Modzelewski, the MM5 by Yongmei Zhou and George Hicks II, and the MC2 by Xingxiu Deng and Yan Shen. The authors thank the three anonymous reviewers for their suggestions to improve the paper.

APPENDIX

Equations for Meteorological Statistical Analysis

Given f_k as forecast precipitation value of the k th forecast, y_k as the corresponding observed value, \bar{f} as mean forecast value, \bar{y} as mean observed value, and N as the number of forecast–observation pairs, then the definitions in the following sections apply:

a. Degree of mass balance (DMB) (Grubišić et al. 2005)

$$\text{DMB} = \frac{\sum_{k=1}^N f_k}{\sum_{k=1}^N y_k}. \quad (\text{A1})$$

DMB is the ratio of the predicted to the observed net water mass over the study period. Values of $\text{DMB} < 1$ indicate that 24-h precipitation is underforecast by the models. A value of $\text{DMB} \approx 1$ indicates the 24-h precipitation forecasts are in balance with the observations.

b. Mean error (ME)

$$\text{ME} = \bar{f} - \bar{y}. \quad (\text{A2})$$

c. Mean absolute error (MAE)

$$\text{MAE} = \frac{1}{N} \sum_{k=1}^N |f_k - y_k|. \quad (\text{A3})$$

d. Mean square error

$$\text{MSE} = \frac{1}{N} \sum_{k=1}^N (f_k - y_k)^2. \quad (\text{A4})$$

e. Root-mean-square error (RMSE)

$$\text{RMSE} = \sqrt{\text{MSE}}. \quad (\text{A5})$$

f. Pearson correlation r

$$r = \frac{\sum_{k=1}^N (f_k - \bar{f})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^N (f_k - \bar{f})^2 \sum_{k=1}^N (y_k - \bar{y})^2}}. \quad (\text{A6})$$

g. Linear error in probability space

Linear error in probability space (LEPS) is defined as the mean absolute difference between the cumulative frequency of the forecasts and the cumulative frequency of the observations (Déqué 2003). LEPS ensures that error in the center of the distribution is treated with more importance than error found in the extreme tail of the distribution. A LEPS skill score can be defined with the climatological median as a reference (Wilks 1995).

h. Brier score (reliability and resolution)

The most widely used EPS evaluation score is the Brier score (Brier 1950; Atger 2003; Gallus et al. 2007), designed to quantify the performance of a probabilistic forecast of a dichotomous event. The Brier score (BS) is defined as the mean square error of the probability forecast,

$$\text{BS} = \frac{1}{N} \sum_{k=1}^N (p_k - o_k)^2, \quad (\text{A7})$$

where N is the number of forecasts, p_k is the probability that forecast precipitation will exceed a given threshold (estimated by the number of ensemble members that exceed that threshold), and o_k is the verifying observation (equal to 1 if the observed precipitation exceeds the threshold, 0 if it does not). The Brier score is negatively oriented, in that a score of 0 is a perfect forecast and increasing Brier score values, to a maximum of 1, indicate deteriorating performance.

Murphy (1973) showed how Eq. (A7) could be partitioned into three parts, measuring the degree of reliability, resolution, and uncertainty in the forecasts and associated observations. In Murphy's decomposition, the verification dataset contains J different probability forecast values, where n_j is the number of cases in the j th forecast category. For each forecast category j , the average of the observations in that category is determined as

$$\bar{o}_j = \frac{1}{n_j} \sum_{k \in n_j} o_k. \quad (\text{A8})$$

Additionally, the overall sample climatology of the defined event is measured as the average of all observations:

$$\bar{o} = \frac{1}{N} \sum_{k=1}^N o_k. \quad (\text{A9})$$

Given definitions (A8) and (A9), Eq. (A7) can be rewritten as

$$\text{BS} = \frac{1}{N} \sum_{j=1}^J n_j (p_j - \bar{o}_j)^2 - \frac{1}{N} \sum_{j=1}^J n_j (\bar{o}_j - \bar{o})^2 + \bar{o}(1 - \bar{o}), \quad (\text{A10})$$

where the first term is the reliability component. Reliability is the correspondence between a given probability and the observed frequency of an event in those cases when the event is forecast with the given probability. The reliability term quantifies the information provided in a reliability diagram.

The second term is the resolution component. The

resolution term indicates the extent that the different forecast categories do in fact reflect different frequencies of occurrence of the observed event.

The last term in the decomposition is the uncertainty. The uncertainty term denotes the intrinsic difficulty in forecasting the event but depends on the observations only, not on the forecasting system.

The Brier score can be converted to a positively oriented skill score,

$$\text{BSS} = 1 - \frac{\text{BS}}{\text{BS}_{\text{ref}}}. \quad (\text{A11})$$

The reference system is often taken to be the low-skill climatological forecast in which the probability of the event is equal to \bar{o} for all forecasts. The Brier score for such climatological forecasts is $\text{BS}_{\text{ref}} = \text{BS}_c = \bar{o}(1 - \bar{o})$.

The Brier skill score is then expressed as

$$\begin{aligned} \text{BSS} &= \frac{\text{resolution}}{\text{uncertainty}} - \frac{\text{reliability}}{\text{uncertainty}} \\ &= \text{relative resolution} - \text{relative reliability}. \end{aligned}$$

A perfect forecast would have the relative resolution equal to 1 and the relative reliability equal to 0.

An adjustment factor must be included in the calculations for Brier skill score to account for SREF systems with different ensemble sizes. Richardson (2001) provides a method of compensating for the effect of ensemble size on the Brier skill score. The relationship between BSS_M (the Brier skill score for an EPS represented by M members) and BSS_∞ (the Brier skill score for the full underlying probability distribution of forecasts as $M \rightarrow \infty$) is

$$\text{BSS}_\infty = \frac{M \cdot \text{BSS}_M + 1}{M + 1}. \quad (\text{A12})$$

The Brier skill score BSS_∞ can then be expressed as a sum of the relative resolution component and the relative reliability component of the Brier score decomposition:

$$\text{BSS}_\infty = \frac{M \cdot \text{BSS}_{M_{\text{relres}}} + 1}{M + 1} - \frac{M \cdot \text{BSS}_{M_{\text{reliab}}}}{M + 1}. \quad (\text{A13})$$

i. Reliability


Reliability is an essential attribute of the quality of probabilistic forecasts (Atger 2003). Reliability is often represented with the aid of a reliability diagram of observed relative frequency o_k versus forecast probability p_k (Wilks 1995; Toth et al. 2003; Clark et al. 2004; Gallus et al. 2007). Ensemble forecasts are converted to

probabilistic forecasts by determining what percentage of the ensemble members meets the specific event criterion. In the current study, the event criterion is that 24-h precipitation exceeds a threshold value of 5, 10, 25, or 50 mm.

Ideally, the probabilistic forecast–observation points lie on the diagonal of the reliability diagram, indicating the event is always forecast at the same frequency it is observed. The reliability component of the Brier score in a graphical representation is the weighted, averaged, squared distance between the reliability curve and the 45° diagonal line. If the points lie above (below) the diagonal, it means the event is underforecast (overforecast). Reliability curves with a zigzag shape centered on the diagonal indicate good reliability represented by a small sample size. Poor reliability can be improved substantially by appropriate a posteriori calibration and/or postprocessing of forecasts delivered from an established system, though it is a difficult task to achieve in a real-time operational EPS (Atger 2003).

Associated with reliability is sharpness, which characterizes the relative frequency of occurrence of the forecast probabilities. Sharpness is often depicted in a histogram indicating the relative occurrence of each forecast probability category. If forecast probabilities are frequently near 0 or near 1, then the forecasts are sharp, indicating the forecasts deviate significantly from the climatological mean, a positive attribute of an ensemble forecast system. Sharpness measures the variability of the forecasts alone, without regard to their corresponding observations; hence, it is not a verification measure in itself. In a perfectly reliable forecast system, sharpness is identical to resolution.

j. Resolution

 A useful probabilistic forecast system must be able to a priori differentiate future weather outcomes, so that differing forecasts are, in fact, associated with distinct verifying observations. This is the most important attribute of a forecast system (Toth et al. 2003) and is called *resolution*.

Resolution cannot be improved through simple adjustment of probability values or statistical postprocessing. Resolution can be gained only by improving the actual forecast model engine that produces the forecasts.

k. ROC curve

Discrimination, which is the converse of resolution, reflects an EPS's ability to distinguish between the occurrence and nonoccurrence of forecast events, in other words, the sensitivity of the probability that an event


TABLE A1. Contingency table for hit rate and false alarm rate calculations. The counts a , b , c , and d are the number of events in each category, out of N total events.

	Observed	Not observed
Forecast	a	b
Not forecast	c	d

was forecast conditioned on whether or not the event was observed. In the case that observed frequencies of forecast events monotonically increase with increasing forecast probabilities, resolution and discrimination convey the same information about an EPS (Buizza et al. 2005). Resolution and discrimination are based on two different factorizations of the forecast–observed pair of events, as described in the distributions-oriented approach to Murphy and Winkler's (1987) general framework for forecast verification.

The relative operating characteristic (ROC) is a verification measure based on signal detection theory that offers a way to examine the discrimination of an EPS (Mason 1982; Mason and Graham 1999; Gallus et al. 2007). The ROC curve is a graph of the hit rate versus the false alarm rate (see below for definitions of these terms and Table A1 for the contingency table basis for this analysis). The ROC measure is based on stratification by observations and, therefore, is independent of reliability and instead provides a direct measure of resolution. The ROC measure is particularly valuable in assessing the general issue of ensemble size–configuration versus model resolution. Additionally, the potential economic value for the cost–loss decision model (Murphy 1977; Katz and Murphy 1997; Richardson 2000; Zhu et al. 2002; Richardson 2003) is uniquely determined from ROC curves (see Part II for an economic analysis of the probabilistic forecasts).

l. Contingency table analysis equations

Given event counts a , b , c , and d from Table A1, the hit rate (H) and false alarm rate (F) are defined as 

$$H = \frac{a}{a + c} \quad \text{and} \quad (\text{A14})$$

$$F = \frac{b}{b + d} \quad (\text{A15})$$

Perfect discrimination is represented by a ROC curve that rises from (0, 0) along the y axis to (0, 1) then straight to (1, 1). The diagonal line represents zero skill, meaning the forecasts show no discrimination among events. The area under the ROC curve (ROC_{Area}) is a measure of skill that can be used to compare different

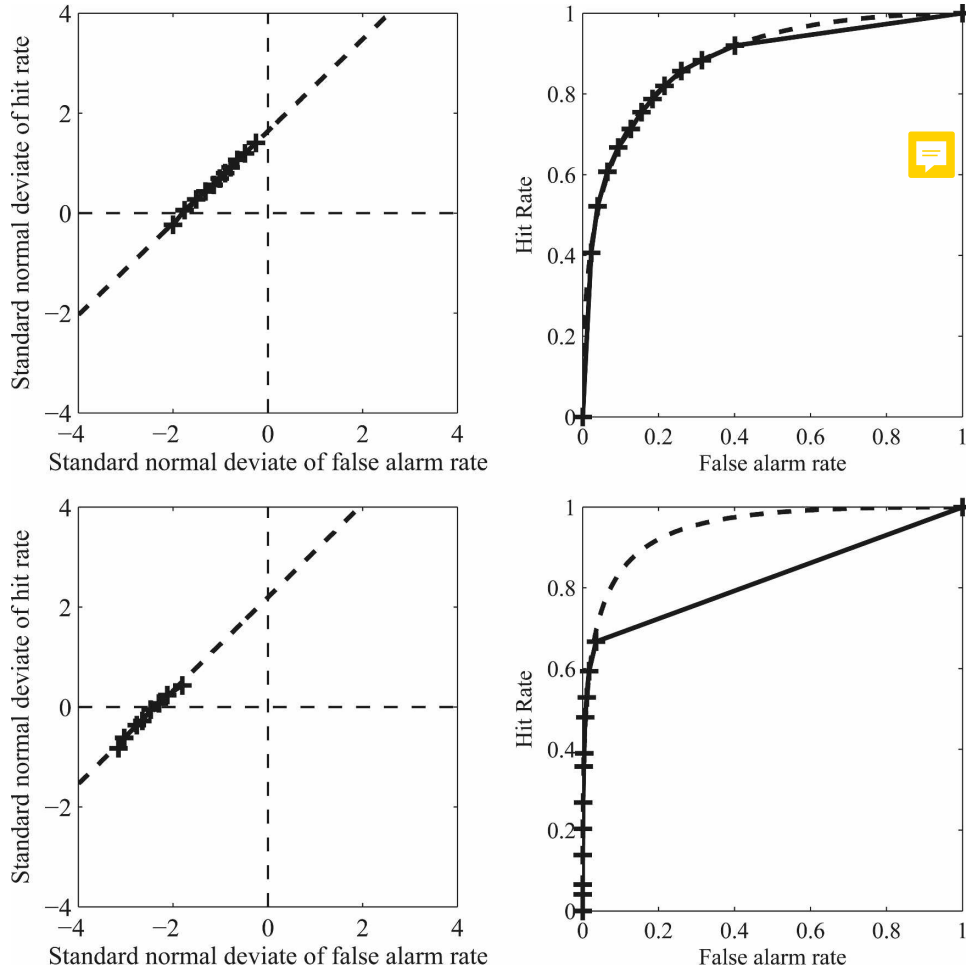


FIG. A1. (left) ROC values transformed to standard normal deviates of hit rate H and false alarm rate F . Crosses are the actual transformations, while the dashed line is the least squares straight-line fit. (right) ROC curves transformed back to the probability axes. Original ROC points are shown by crosses connected together by a solid line. Transforming the straight line from the plots on the left back to probability axes gives the dashed line in the plots on the right. The area under the transformed, or modeled, ROC curve is A_z . (top) Day-1 full 11-member ensemble forecasts with the 5 mm day⁻¹ precipitation threshold. (bottom) The same ensemble as in the top panels but for the more rare 50 mm day⁻¹ precipitation threshold.

probabilistic forecast systems. Perfect discrimination results in a ROC area value of 1.0 while the no-skill diagonal corresponds to a ROC area value of 0.5. From these values, a ROC skill score (ROCSS) can be defined to range from 0 (no-skill forecasts) to 1 (perfectly discriminating forecasts):

$$\text{ROCSS} = 2\text{ROC}_{\text{Area}} - 1, \quad (\text{A16})$$

The ROC area is determined from A_z , the area under the *modeled* ROC on probability axes (Mason 1982; Swets and Pickett 1982; Swets 1988; Mason 2003). In practice, the ROC curve is transformed (see Figs. A1

and A2) to a plot on normal deviate axes that results in a straight line [under the assumption of a normal (Gaussian) probability distribution (Swets and Pickett 1982)]. The straight line can be estimated from the data using a least squares method of linear interpolation. The slope and y-axis intercept of this line can be used to evaluate $z(A)$:

$$z(A) = \frac{s\Delta m}{(1 + s^2)^{1/2}}, \quad (\text{A17})$$

where s is the slope of the line and Δm is the y-axis intercept (with the sign reversed). Transforming $z(A)$ back to probability space through the use of the cumu-

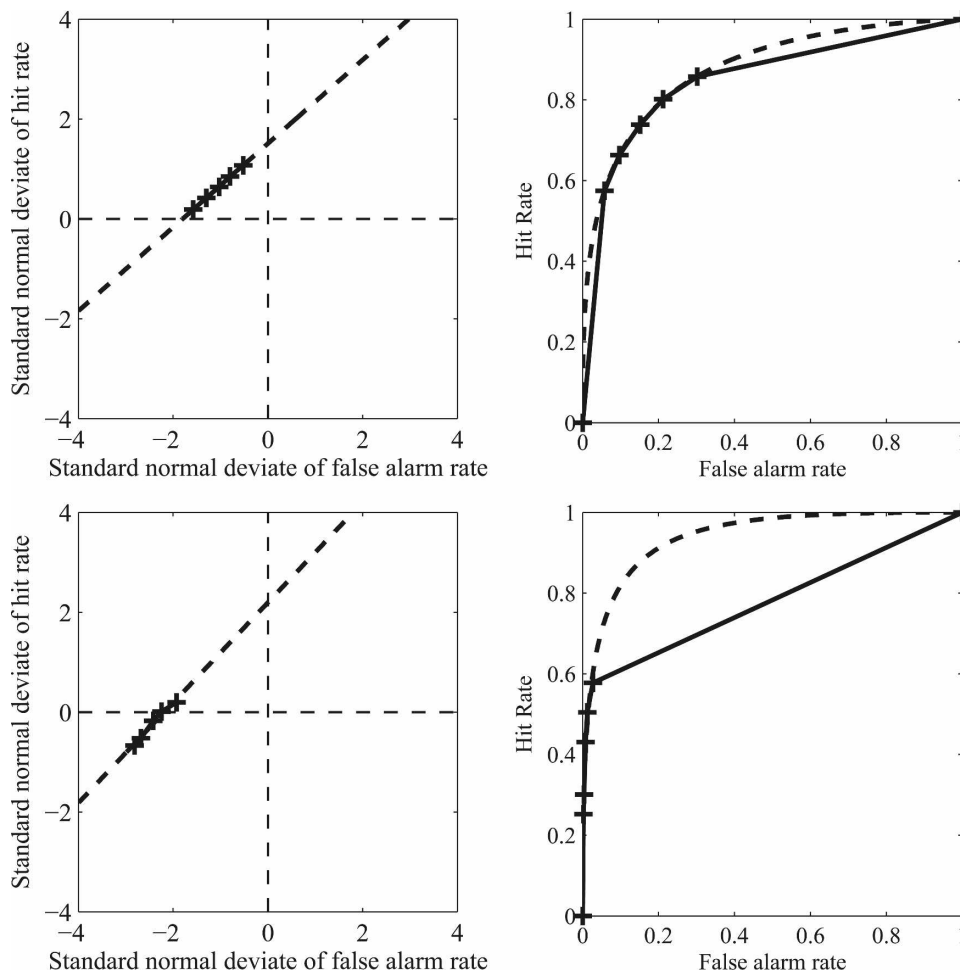


FIG. A2. Same as in Fig. A2, but for the highest-resolution ensemble (vhires5).

lative standardized normal distribution gives A_z as the area under the ROC curve on probability axes.

m. Equal likelihood

Rank histograms, also known as Talagrand diagrams (Anderson 1996; Talagrand et al. 1998), provide a necessary but not sufficient test for evaluating whether the forecast and verification are sampled from the same probability distribution. The rank histogram is a useful measure of the realism of an ensemble forecast system (Hou et al. 2001), providing a measure of reliability (Candille and Talagrand 2005). For an ensemble forecast with M members, there will be $M + 1$ intervals defined by the members, including the two open-ended intervals. A rank histogram is built by accumulating the number of cases that an observation falls in each of the $M + 1$ intervals.

In an ideal ensemble, the range of the ensemble forecast members represents the full spread of the prob-

ability distribution of observations, and each member exhibits an equal likelihood of occurrence—the resulting rank histogram is flat. Ensemble forecasts with insufficient spread to adequately represent the probability distribution of observations are indicated by a U-shaped histogram. Rank histograms computed from operational ensembles are commonly U shaped, which is traditionally interpreted as the consequence of the lack of ensemble spread in the EPS (Anderson 1996; Atger 2003). Alternately, U-shaped rank histograms may be interpreted as a consequence of conditional model biases (Hamill 2001), rather than pointing to a weakness of the method used for generating the ensemble.

A measure of the deviation of the rank histogram from flatness is described by Candille and Talagrand (2005) for an EPS with M members and N available forecast–observation pairs. The number of values in the i th interval of the histogram is given by s_i . For a reliable

system in which the histogram is flat, $s_i = N/(M + 1)$ for each interval i . The quantity

$$\Delta = \sum_{i=1}^{M+1} \left(s_i - \frac{N}{M+1} \right)^2 \quad (\text{A18})$$

measures the deviation of the histogram from flatness. For a reliable system, the base value is $\Delta_0 = NM/(M + 1)$. The ratio

$$\delta = \Delta/\Delta_0 \quad (\text{A19})$$

is evaluated as the overall measure of the flatness of an ensemble prediction system rank histogram. A value of δ that is significantly larger than 1 indicates the system does not reflect equal likelihood of ensemble members.

An alternative summary index based on the rank histogram, called a reliability index (RI), was recently introduced by Delle Monache (2005); see also Delle Monache et al. (2006). The RI measures the variation of the rank histogram from its ideal “flat” shape and is normalized so that ensembles of different sizes can be compared with each other.

REFERENCES

- Anderson, J. L., 1996: A method for producing and evaluating probabilistic precipitation forecasts from ensemble model integrations. *J. Climate*, **9**, 1518–1529.
- Atger, F., 2003: Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: Consequences for calibration. *Mon. Wea. Rev.*, **131**, 1509–1523.
- Benoit, R., M. Desgagne, P. Pellerin, S. Pellerin, and Y. Chartier, 1997: The Canadian MC2: A semi-Lagrangian, semi-implicit wideband atmospheric model suited for finescale process studies and simulation. *Mon. Wea. Rev.*, **125**, 2383–2415.
- Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3.
- Bright, D. R., M. S. Wandishin, R. E. Jewell, and S. J. Weiss, 2005: A physically based parameter for lightning prediction and its calibration in ensemble forecasts. Preprints, *Conf. on Meteor. Applications of Lightning Data*, San Diego, CA, Amer. Meteor. Soc., 4.3. [Available online at <http://ams.confex.com/ams/pdfpapers/84173.pdf>.]
- Buizza, R., A. Hollingsworth, F. Lalaurette, and A. Ghelli, 1999: Probabilistic predictions of precipitation using the ECMWF Ensemble Prediction System. *Wea. Forecasting*, **14**, 168–189.
- , P. L. Houtekamer, Z. Toth, G. Pellerin, M. Wei, and Y. Zhu, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Wea. Rev.*, **133**, 1076–1097.
- Candille, G., and O. Talagrand, 2005: Evaluation of probabilistic prediction systems for a scalar variable. *Quart. J. Roy. Meteor. Soc.*, **131**, 2131–2150.
- Cheng, W. Y. Y., and W. J. Steenburgh, 2007: Strengths and weaknesses of MOS, running-mean bias removal, and Kalman filter techniques for improving model forecasts over the western United States. *Wea. Forecasting*, **22**, 1304–1318.
- Clark, M., S. Gangopadhyay, L. Hay, B. Rajagopalan, and R. Wilby, 2004: The Schaake shuffle: A method for reconstructing space–time variability in forecasted precipitation and temperature fields. *J. Hydrometeorol.*, **5**, 243–262.
- Davis, C., and F. Carr, 2000: Summary of the 1998 Workshop on Mesoscale Model Verification. *Bull. Amer. Meteor. Soc.*, **81**, 809–819.
- Delle Monache, L., 2005: Ensemble-averaged, probabilistic, and Kalman-filtered regional ozone forecasts. Ph.D. thesis, University of British Columbia, 211 pp.
- , J. P. Hacker, Y. Zhou, X. Deng, and R. B. Stull, 2006: Probabilistic aspects of meteorological and ozone regional ensemble forecasts. *J. Geophys. Res.*, **111**, D24307, doi:10.1029/2005JD006917.
- Déqué, M., 2003: Continuous variables. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, I. Jolliffe and D. B. Stephenson, Eds., Wiley, 97–119.
- Du, J., S. Mullen, and F. Sanders, 1997: Short-range ensemble forecasting of quantitative precipitation. *Mon. Wea. Rev.*, **125**, 2427–2459.
- Dudhia, J., J. Klemp, W. C. Skamarock, D. Dempsey, Z. I. Janjić, S. G. Benjamin, and J. M. Brown, 1998: A collaborative effort towards a future community mesoscale model. Preprints, *12th Conf. on Numerical Weather Prediction*, Phoenix, AZ, Amer. Meteor. Soc., 42–43.
- Eckel, F. A., and C. F. Mass, 2005: Aspects of effective mesoscale, short-range ensemble forecasting. *Wea. Forecasting*, **20**, 328–350.
- Gallus, W. A., Jr., M. E. Baldwin, and K. L. Elmore, 2007: Evaluation of probabilistic precipitation forecasts determined from Eta and AVN forecasted amounts. *Wea. Forecasting*, **22**, 207–215.
- Grell, G., J. Dudhia, and D. R. Stauffer, 1994: A description of the fifth-generation Penn State/NCAR Mesoscale Model (MM5). NCAR Tech. Rep. TN-398+STR, 121 pp.
- Grubišić, V., R. K. Vellore, and A. W. Huggins, 2005: Quantitative precipitation forecasting of wintertime storms in the Sierra Nevada: Sensitivity to the microphysical parameterization and horizontal resolution. *Mon. Wea. Rev.*, **133**, 2834–2859.
- Hacker, J. P., E. S. Krayenhoff, and R. B. Stull, 2003: Ensemble experiments on numerical weather prediction error and uncertainty for a North Pacific forecast failure. *Wea. Forecasting*, **18**, 12–31.
- Hamill, T. M., 1997: Reliability diagrams for multicategory probabilistic forecasts. *Wea. Forecasting*, **12**, 736–741.
- , 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560.
- , S. L. Mullen, C. Snyder, Z. Toth, and D. P. Baumhefner, 2000: Ensemble forecasting in the short to medium range: Report from a workshop. *Bull. Amer. Meteor. Soc.*, **81**, 2653–2664.
- Hou, D., E. Kalnay, and K. K. Droegemeier, 2001: Objective verification of the SAMEX '98 ensemble forecasts. *Mon. Wea. Rev.*, **129**, 73–91.
- Jolliffe, I. T., and D. B. Stephenson, 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. J. Wiley, 254 pp.
- Jones, M. S., B. A. Colle, and J. S. Tongue, 2007: Evaluation of a mesoscale short-range ensemble forecast system over the northeast United States. *Wea. Forecasting*, **22**, 36–55.
- Katz, R. W., and A. H. Murphy, 1997: *Economic Value of Weather and Climate Forecasts*. Cambridge University Press, 222 pp.
- Krzysztofowicz, R., W. J. Drzal, T. R. Drake, J. C. Weyman, and L. A. Giordano, 1993: Probabilistic quantitative precipitation forecasts for river basins. *Wea. Forecasting*, **8**, 424–439.

- Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.*, **30**, 291–303.
- , 2003: Binary events. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, I. Jolliffe and D. B. Stephenson, Eds., Wiley, 137–163.
- , and N. E. Graham, 1999: Conditional probabilities, relative operating characteristics, and relative operating levels. *Wea. Forecasting*, **14**, 713–725.
- Mass, C. F., D. Owens, K. Westrick, and B. Cole, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bull. Amer. Meteor. Soc.*, **83**, 407–430.
- McCollor, D., and R. Stull, 2008a: Hydrometeorological accuracy enhancement via postprocessing of numerical weather forecasts in complex terrain. *Wea. Forecasting*, **23**, 131–144.
- , and —, 2008b: Hydrometeorological short-range ensemble forecasts in complex terrain. Part II: Economic evaluation. *Wea. Forecasting*, **23**, 557–574.
- McMurdie, L., and C. Mass, 2004: Major numerical forecast failures over the northeast Pacific. *Wea. Forecasting*, **19**, 338–356.
- Mullen, S. L., and R. Buizza, 2002: The impact of horizontal resolution and ensemble size on probabilistic forecasts of precipitation by the ECMWF Ensemble Prediction System. *Wea. Forecasting*, **17**, 173–191.
- Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595–600.
- , 1977: The value of climatological, categorical, and probabilistic forecasts in the cost-loss ratio situation. *Mon. Wea. Rev.*, **105**, 803–816.
- , 1991: Forecast verification: Its complexity and dimensionality. *Mon. Wea. Rev.*, **119**, 1590–1601.
- , and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.*, **115**, 1330–1338.
- Nutter, P., D. Stensrud, and M. Xue, 2004: Effects of coarsely resolved and temporally interpolated lateral boundary conditions on the dispersion of limited-area ensemble forecasts. *Mon. Wea. Rev.*, **132**, 2358–2377.
- Richardson, D. S., 2000: Skill and relative economic value of the ECMWF Ensemble Prediction System. *Quart. J. Roy. Meteor. Soc.*, **126**, 649–667.
- , 2001: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quart. J. Roy. Meteor. Soc.*, **127**, 2473–2489.
- , 2003: Economic value and skill. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, I. Jolliffe and D. B. Stephenson, Eds., Wiley, 164–187.
- Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang, and J. G. Powers, 2005: A description of the Advanced Research WRF version 2. NCAR Tech. Rep. TN-468+STF, 88 pp.
- Spagnol, J., C. Readyhough, M. Stull, J. Mundy, R. Stull, S. Green, and G. Schajer, 2004: Rocketsonde buoy system observing system simulation experiments. Preprints, *20th Conf. on Weather Analysis and Forecasting/16th Conf. on Numerical Weather Prediction/Symp. on Forecasting the Weather and Climate of the Atmosphere and Ocean*, Seattle, WA, Amer. Meteor. Soc., J7.7. [Available online at <http://ams.confex.com/ams/pdfpapers/68557.pdf>.]
- Stensrud, D. J., and N. Yussouf, 2007: Reliable probabilistic quantitative precipitation forecasts from a short-range ensemble forecasting system. *Wea. Forecasting*, **22**, 3–17.
- , J. Bao, and T. T. Warner, 2000: Using initial conditions and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. *Mon. Wea. Rev.*, **128**, 2077–2107.
- Swets, J. A., 1988: Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293.
- , and R. M. Pickett, 1982: *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. Academic Press, 253 pp.
- Talagrand, O., R. Vautard, and B. Strauss, 1998: Evaluation of probabilistic prediction systems. *Proc. Workshop on Predictability*, Reading, United Kingdom, ECMWF, 1–25.
- Toth, Z., O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, I. Jolliffe and D. B. Stephenson, Eds., Wiley, 137–163.
- Wandishin, M. S., S. L. Mullen, D. J. Stensrud, and H. E. Brooks, 2001: Evaluation of a short-range multimodel ensemble system. *Mon. Wea. Rev.*, **129**, 729–747.
- Weber, F., L. Perreault, and V. Fortin, 2006: Measuring the performance of hydrological forecasts for hydropower production at BC Hydro and Hydro-Québec. Preprints, *18th Conf. on Climate Variability and Change*, Atlanta, GA, Amer. Meteor. Soc., 8.5. [Available online at <http://ams.confex.com/ams/pdfpapers/101960.pdf>.]
- Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction*. Academic Press, 467 pp.
- Wilson, L. J., 2000: Comments on “Probabilistic predictions of precipitation using the ECMWF Ensemble Prediction System.” *Wea. Forecasting*, **15**, 361–364.
- Yuan, H., S. Mullen, X. Gao, S. Sorooshian, J. Du, and H.-M. H. Juang, 2005: Verification of probabilistic quantitative precipitation forecasts over the southwest United States during winter 2002/03 by the RSM ensemble system. *Mon. Wea. Rev.*, **133**, 279–294.
- Yussouf, N., and D. J. Stensrud, 2007: Bias-corrected short-range ensemble forecasts of near-surface variables during the 2005/06 cool season. *Wea. Forecasting*, **22**, 1274–1286.
- Zhu, Y., Z. Toth, R. Wobus, D. Richardson, and K. Mylne, 2002: The economic value of ensemble-based weather forecasts. *Bull. Amer. Meteor. Soc.*, **83**, 73–83.