Verification of 24-h Quantitative Precipitation Forecasts over the Pacific Northwest from a High-Resolution Ensemble Kalman Filter System

PHILLIPA COOKSON-HILLS, DANIEL J. KIRSHBAUM, MADALINA SURCEL, AND JONATHAN G. DOYLE

Department of Atmospheric and Oceanic Sciences, McGill University, Montréal, Québec, Canada

LUC FILLION, DOMINIK JACQUES, AND SEUNG-JONG BAEK

Data Assimilation and Satellite Meteorology Research Section, Environment and Climate Change Canada, Dorval, Québec, Canada

(Manuscript received 14 October 2016, in final form 21 February 2017)

ABSTRACT

Environment and Climate Change Canada (ECCC) has recently developed an experimental high-resolution EnKF (HREnKF) regional ensemble prediction system, which it tested over the Pacific Northwest of North America for the first half of February 2011. The HREnKF has 2.5-km horizontal grid spacing and assimilates surface and upper-air observations every hour. To determine the benefits of the HREnKF over less expensive alternatives, its 24-h quantitative precipitation forecasts are compared with those from a lower-resolution (15 km) regional ensemble Kalman filter (REnKF) system and to ensembles directly downscaled from the REnKF using the same grid as the HREnKF but with no additional data assimilation (DS). The forecasts are verified against rain gauge observations and gridded precipitation analyses, the latter of which are characterized by uncertainties of comparable magnitude to the model forecast errors. Nonetheless, both deterministic and probabilistic verification indicates robust improvements in forecast skill owing to the finer grids of the HREnKF and DS. The HREnKF exhibits a further improvement in performance over the DS in the first few forecast hours, suggesting a modest positive impact of data assimilation. However, this improvement is not statistically significant and may be attributable to other factors.

1. Introduction

The Pacific Northwest (PNW) is a mountainous coastal region encompassing the extreme northwestern United States and the southwestern coast of Canada (Fig. 1a). Its precipitation distribution and type are variable and heavily influenced by its close proximity to the Pacific Ocean and its complex topography. It receives the majority of its precipitation during the wintertime, when the placement of the midlatitude storm track causes it to experience a succession of landfalling midlatitude cyclones (e.g., Mote et al. 2003). The moist maritime flow associated with these storms, when forced to ascend the steep regional terrain, produces heavy precipitation on windward (south to west facing) mountain slopes. As one of the most dramatic rain shadows on earth, the Olympic Mountains in Washington State are home to midlatitude rain forests on their southwestern slopes (up to 4.5 m of annual precipitation) and semiarid conditions on their lee (less than 25 cm of annual precipitation) (e.g., Minder et al. 2008).

As a populous region with major urban centers (e.g., Portland, Oregon; Seattle, Washington; and Vancouver, British Columbia, Canada), the PNW is strongly affected by heavy precipitation events. Accurate precipitation forecasts of such events are thus essential to mitigating the risks associated with flash flooding, landslides, and other natural hazards. However, accurate forecasting is challenged by the scarcity of observations over and upstream of the region, the small-scale variability imparted by the terrain, and the complex mixed-phase microphysical processes in orographic clouds, among other things (e.g., Mass and Ferber 1990; Lin et al. 2000; Mass et al. 2002). To maximize forecast skill in this region (and in general), model forecasts require 1) an accurate initialization, 2) an explicit consideration of model uncertainties, and 3) a sufficiently
fine grid to capture key physical processes regulating precipitation. The innovation and computer power required to simultaneously achieve these three ends renders it a new frontier in numerical weather prediction (e.g., Clark et al. 2016).

Accurate initialization requires modern data assimilation systems, which typically use variational methods or ensemble Kalman filters (EnKFs) (Houtekamer et al. 2005; Whitaker et al. 2008; Zhang et al. 2011). The key advantage of EnKFs is their dual use of forecast ensembles for 1) data assimilation, to generate flow-dependent background covariance matrices that capture the “errors of the day,” and 2) ensemble prediction (e.g., Ehrendorfer and Tribbia 1997; Yang et al. 2012). Ensemble forecasts sample the probability distributions of uncertain initial conditions and/or model physics schemes to provide a set of plausible future realizations. Because of their potential for improving forecast skill and quantifying forecast reliability and uncertainty (e.g., Kalnay 2003), ensembles have become nearly universal in global numerical weather prediction (NWP) and are gaining popularity in high-resolution regional NWP (e.g., Clark et al. 2012, 2016; Barrett et al. 2016).

As computer power increases, regional forecast models are increasingly using “convective scale” O(1) km grids (e.g., Clark et al. 2012; Barrett et al. 2016). Because these grids explicitly represent the embedded moist instabilities that often develop in orographic clouds (e.g., Houze and Medina 2005; Kirshbaum et al. 2007), they may eliminate systematic biases in orographic precipitation linked to cumulus parameterization (Schwitalla et al. 2008). However, higher grid resolution does not always translate to improved forecast skill over the PNW (e.g., Colle and Mass 2000; Mass et al. 2002). This counterintuitive finding may arise in part from the “double penalty” problem inherent to convective-scale point verification, where a slight displacement of a small feature (e.g., a convective cell) can incur twice the error of a null forecast (Gilleland et al. 2009).

To combine the benefits of the EnKF with those of finer grid resolution, Environment and Climate Change Canada (ECCC) recently tested an experimental high-resolution ensemble Kalman filter (HREnKF) system on a 2.5-km grid over the PNW during a portion of the 2011 rainy season (Jacques et al. 2017). Herein we evaluate and compare quantitative precipitation forecasts (QPFs) from the HREnKF against those from two less computationally expensive ensemble systems, the 15-km continental-scale regional EnKF (REnKF; Baek et al. 2012) and an ensemble downscaled from the REnKF using the same 2.5-km grid as the HREnKF but with no high-resolution data assimilation (DS) (the latter described by Jacques et al. 2017). While the HREnKF is expected to exhibit the highest skill of all three ensembles, its performance is still limited by the availability of observations, errors inherited from its parent ensemble, and uncertainties in its representation of small-scale processes. In their verification of the HREnKF using surface observations, Jacques et al. (2017) did not find an obvious improvement relative to the DS. However, their study did not include a formal QPF verification, nor did it investigate the benefit of the HREnKF over the lower-resolution REnKF.

A variety of metrics have been devised for QPF verification, which, like the forecasts themselves, can be classified as deterministic or probabilistic. Because each metric gives a different perspective on forecast skill (e.g., Casati et al. 2008; Bouallgue and Theis 2014;
Wilks 2011), a combination of metrics is often useful to avoid misinterpretation. To overcome the aforementioned double-penalty problem, metrics have been developed to verify the full range of spatial scales in the precipitation field rather than just point locations (e.g., Gilleland et al. 2009). However, these methods require a reliable gridded precipitation analysis, which is not always available.

The complex terrain of the PNW renders it a challenging place for QPF verification (e.g., Westrick et al. 1999). Rain and snow gauges are sparse in rural and mountainous regions, and gauges are often placed in easily accessible valleys that are unrepresentative of the surrounding terrain (e.g., Wood et al. 2000). Gauges can also exhibit snow undercatch in strong winds, which may be accounted for in postprocessing if coincident wind and temperature data are available (Yang et al. 1995). While radar can provide continuous, high-resolution precipitation measurements, its data coverage and usefulness are limited by beam blocking over mountains and brightband effects (e.g., Austin 1987; Joss and Lee 1995; Fabry and Zawadzki 1995). A composite of the lowest-elevation scan of the regional U.S. and Canadian operational radars during February 2011 reveals widespread beam blockage and areas where the beam is located several kilometers above the surface (Fig. 1b).

Motivated by the need for reliable forecasts of heavy precipitation events over the PNW, this study has three objectives: 1) to evaluate the uncertainty of precipitation verification products over the mountainous PNW, 2) to thoroughly quantify HREnKF QPF skill in such events, and 3) to quantify any related benefits of the HREnKF over less expensive alternatives. The model setup is described in section 2, with section 3 presenting the observational products, and section 4 providing the verification metrics. Section 5 presents the model verification, followed by a discussion (section 6) and conclusions (section 7).

2. Model setup

The model chain used for the ensemble simulations is detailed in Jacques et al. (2017) and thus only briefly reviewed here. Using the Global Environmental Multiscale (GEM) operational forecast model (Girard et al. 2014), ECCC runs an operational global EnKF (GEnKF) with ~50-km grid spacing and 256 members (Houtekamer et al. 2014), which drives an experimental 15-km, 256-member limited-area regional EnKF over North America (Baek et al. 2012). The REnKF, in turn, drives the 2.5-km, 96-member HREnKF over the PNW (Fig. 2). All three systems use continuous cycling. The HREnKF domain covers the PNW and extends westward over the open ocean, allowing it to drive regional ocean models used by Canada’s Marine Environmental Observation Prediction and Response (MEOPAR) research network.1

Both the GEnKF and REnKF assimilate surface, aircraft, radiosonde, GPS radio occultation(s), and satellite winds and radiances. The HREnKF assimilates the same observations as the REnKF (except for satellite radiances) but differs in the frequency of the analysis.

---

1 More information on MEOPAR is available online (http://meopar.ca/).
(hourly in the HREnKF vs 6 hourly in REnKF) and in the use of a smaller horizontal localization radius (60–100 km in the HREnKF vs 2100–3000 km in the REnKF). The finer grid and more frequent cycling of the HREnKF render it more capable of capturing high-frequency mesoscale processes than its parent models. However, this frequent cycling is very expensive computationally, which, given available computing resources, limited the experimental period to 14 days (1–14 February 2011, which consumed 6 months’ worth of wall-clock time). February 2011 was chosen because it serves as a standard test bed for ECCC model evaluation and coincides with the PNW wet season.

Other notable differences between the HREnKF and REnKF include explicit representation of convection in the former but parameterized convection in the latter using the Kain–Fritsch scheme (Kain and Fritsch 1993). The HREnKF also uses prognostic two-moment microphysics (Milbrandt and Yau 2005) rather than diagnostic microphysics (Sundqvist 1978). Moreover, whereas the REnKF uses a “cold” startup in which all variables not analyzed by the EnKF (including hydrometeors) are reset to zero at the beginning of each simulation, the HREnKF conserves all such fields during the analysis. Thus, the HREnKF minimizes errors associated with precipitation regeneration during model spinup.

Our verification focuses on 24-h forecasts of heavy PNW precipitation, although some shorter-range evaluation is also included. This integration period was chosen because of the increased availability of gauge data at 24-h intervals than at smaller time intervals. However, because the marginal benefits of convective-scale data assimilation on limited-area grids tend to fade rapidly in the first few forecast hours (e.g., Kain et al. 2010; Surcel et al. 2015), these benefits of the HREnKF may be obscured in 24-h forecasts. Nevertheless, 24-h QPFs are important for end users who require sufficient lead time to minimize weather-related risks. The forecasts are launched at 0000 UTC on every day of the study period where the maximum gauge-observed precipitation over the PNW exceeded 40 mm (3, 4, 6, and 11–14 February 2011). We caution that because the focus is restricted to heavier, more strongly forced precipitation events, our findings may not carry over to lighter, weakly forced events. Moreover, our small sampling of events (again, constrained by the HREnKF’s computational expense) limits the robustness of our findings.

HREnKF and REnKF analyses are used to drive 40 ensemble forecasts each day. Because these two ensembles differ primarily in their grid resolution, comparisons between them mainly isolate this factor. To also investigate the impacts of HREnKF data assimilation, a third set of ensemble forecasts, downscaled from the 15-km REnKF to the 2.5-km HREnKF grid with no high-resolution data assimilation (called the DS, or downscaled, ensemble), is conducted. Because the DS ensemble is initialized from REnKF analyses, it also uses a cold start.

3. Observations

Table 1 shows the average and maximum precipitation (mm) for all seven events based on rain gauge and radar data (using the observational networks discussed below). Because 14 February 2011 received the most precipitation of all events, we select it as an illustrative case study.

a. Radar

Low-level radar composites over the PNW were provided by ECCC at 1-km grid spacing. These composites include two Canadian C-band radars and two U.S. NEXRAD Weather Surveillance Radar-1988 Doppler (WSR-88D) S-band radars (Fig. 1b). For the former, the lowest-elevation angle plan position indicator (PPI) scans were used, along with a simple de-cluttering scheme using consecutive Doppler velocity scans to detect and remove nonmeteorological targets with highly fluctuating velocities (at ranges below 125 km). The resulting reflectivity maps were remapped onto a common Cartesian regional grid using nearest-neighbor interpolation. Low-level U.S. radar composites were provided by the Warning Decision Support System–Integrated Information (WDSS-II; Lakshmanan et al. 2006). These data were also remapped onto the common grid using nearest-neighbors interpolation. For points with data from more than one radar product, the average reflectivity from the two nearest radars was used.

<table>
<thead>
<tr>
<th>Day</th>
<th>Gauge avg</th>
<th>Gauge max</th>
<th>Radar avg</th>
<th>Radar max</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>10.1</td>
<td>75.1</td>
<td>9.2</td>
<td>119.8</td>
</tr>
<tr>
<td>4</td>
<td>8.0</td>
<td>60.2</td>
<td>15.4</td>
<td>110.6</td>
</tr>
<tr>
<td>6</td>
<td>7.0</td>
<td>42.6</td>
<td>8.5</td>
<td>119.0</td>
</tr>
<tr>
<td>11</td>
<td>6.8</td>
<td>80.8</td>
<td>6.4</td>
<td>117.4</td>
</tr>
<tr>
<td>12</td>
<td>15.0</td>
<td>112.1</td>
<td>12.2</td>
<td>117.8</td>
</tr>
<tr>
<td>13</td>
<td>9.0</td>
<td>93.2</td>
<td>4.5</td>
<td>119.6</td>
</tr>
<tr>
<td>14</td>
<td>17.7</td>
<td>108.3</td>
<td>25.1</td>
<td>116.1</td>
</tr>
</tbody>
</table>
values exceeding 35 dB are uncommon and thus assumed to lie within the radar bright band, for which the relation $Z = 168 R^{1.58}$ is used. Among numerous retrievals under consideration, this combination of $Z-R$ relations best matched the corresponding 24-h rain gauge accumulations over the full study period (not shown). Areas of terrain blocking in the lowest-elevation scan, or where the lowest beam height exceeded 3 km in Fig. 1b, are treated as missing data. The 24-h radar-derived accumulation on 14 February 2011 is shown in Fig. 3a, revealing patches of heavy precipitation to the northwest and east of the Olympics and a rain shadow to the northeast of the Olympics.

b. Precipitation gauges

Daily rain and snow gauge accumulations over the PNW were accessed from the Pacific Climate Impacts Consortium (https://www.pacificclimate.org/) and Mesowest (http://mesowest.utah.edu/) online databases. While most of these gauges were located at elevations below 1000 m where the dominant precipitation type is rain, a handful were located over the higher terrain where large snow undercatch can occur (e.g., Yang et al. 1999). Thus, we implemented a correction to gauges with daily mean temperatures below 0°C, provided that wind speed was also reported (if not, the gauge data were discarded). Because the majority of precipitation gauges used in this study are the Geonor T-200B type, we applied a corresponding correction factor $k = e^{0.2U}$ from Smith (2007), where $U$ is the daily average wind speed (m s$^{-1}$). As undercatch is generally minor at U.S. Snowpack Telemetry (SNOTEL) network stations in the Olympics and Cascades (e.g., Colle and Mass 2000), no correction was applied there. Although the gauge data available online were already quality controlled, we additionally omitted gauges that disagreed with two or more surrounding gauges within a 2-km radius by more than 40 mm. Finally, any accumulations of less than 0 mm or exceeding 120 mm were omitted.

The 24-h gauge accumulations on 14 February 2011 are broadly consistent with the corresponding radar-derived accumulation except over the Puget Sound region, where a patch of precipitation exceeding 60 mm in the radar retrieval is absent in the gauge data (Figs. 3a,b). The average gauge precipitation amount over the same region is only around 15 mm, a discrepancy that is attributable to a quasi-stationary area of >30-dBZ radar reflectivity over the majority of the day. This large reflectivity is likely associated with the radar beam intersecting the bright band. Despite our attempts to account for this effect, the retrieved precipitation is still grossly overestimated. This case exemplifies the large
discrepancies that may arise between radar and rain gauge observations in the PNW and emphasizes that either network alone may be insufficient to verify the model.

The majority of gauges in the PNW are located in three areas: Vancouver Island (VI), the Puget Sound (PS) area, and the Vancouver metropolitan area (VA) (Fig. 3b). Because the topography and precipitation can differ markedly between these regions, each is treated separately for some aspects of the deterministic point verification. Table 2 reports the total numbers of gauges in the PNW, VI, PS, and the VA regions for each event, indicating the highest gauge density in the PS region.

c. Gridded precipitation analyses

Spatial verification generally requires observations on a regular grid, which can be produced using a combination of data sources (gauge, radar, model analyses, etc.). While two such analyses are already routinely available and discussed below, each has disadvantages that limit its utility for this study. Therefore, we also developed a third such analysis. This use of multiple products provides insight into the uncertainty of the precipitation analyses.

1) STAGE IV

The National Centers for Environmental Prediction (NCEP) stage IV precipitation product is a mosaic of regional multisensor analyses (including WSR-88D radar and rain gauges) produced by the National Weather Service (NWS) River Forecast Centers (RFCs) (Lin and Mitchell 2005). Each RFC employs both manual and automatic quality control measures, including accounting for orographic beam blocking, quality controlling near-ground reflectivities, and applying a mean-field radar and rain gauge bias correction or site-specific calibration for each radar (Fulton 2002). Hourly and 6-hourly precipitation accumulations are available on a polar stereographic grid (with a spacing of 4.7625 km at 60°N) from the stage IV website (http://www.emc.ncep.noaa.gov/mmb/lin/pcpanl/stage4/). Four 6-h analyses are summed to obtain 24-h analyses for each day. The resulting field on 14 February 2011 indicates broadly moderate precipitation except for a maximum on the southwest side of the Olympics that locally exceeds 80 mm (Fig. 4a). This distribution is largely consistent with the corresponding gauge data but not with the radar analysis (Fig. 3). Because stage IV uses the most advanced algorithm of any gridded analyses herein, we place the most confidence in it. However, because it is limited to the United States, other analyses are required for verification over southwestern Canada.

2) CaPA

The Canadian Precipitation Analysis (CaPA) is a data assimilation system that produces near-real-time 6-h precipitation analyses at 15-km grid resolution, four times a day (Mahfouf et al. 2007; Fortin and Roy 2011; Fortin et al. 2012). Its background field is obtained from ECCC’s Regional Deterministic Prediction System (RDPS), and observations are incorporated from a variety of sources, including manual and automatic weather stations, aviation routine weather reports (METARs), and multiple cooperative gauge networks (operational radar data were not included in February 2011 but have been added since). As with stage IV, four 6-hourly analyses are summed to obtain 24-h accumulations for each day. The CaPA analysis on 14 February 2011 shows a similar placement of precipitation over the Olympics as the corresponding stage IV analysis, except for a southward shift of the maximum, smaller accumulations, and less mesoscale detail (Fig. 4b).

3) KRIGING WITH EXTERNAL DRIFT (KED)

Given that the stage IV analysis is restricted to the United States, and the CaPA analysis is based on uncertain model forecasts, we seek a third gridded product that relies purely on gauge and radar data over the entire PNW. In an evaluation of various radar–gauge merging methods over the United Kingdom, Goudenhoofdt and Delobbe (2009) found the kriging with external drift (KED) method to perform best over all network densities and seasons, so we choose it herein. This technique is based on ordinary kriging (OK), which uses a network of gauge observations to interpolate precipitation to a uniform grid. The interpolation consists of a weighted average of all observations surrounding the point of interest, where the weights are determined by the spatial correlation of the observations, as quantified through a variogram.

Our OK implementation uses a spherical parametric variogram for computational efficiency, as in Jewell and Gaussiat (2015). The resulting analysis on 14 February 2011, on the same 2.5-km grid as the HREnKF but masking the oceans, is smooth in poorly observed
regions and more detailed in well observed areas like the Puget Sound (Fig. 4c). Precipitation is again maximized over the Olympics, but the maximum is larger than in CaPA and is more focused over the high terrain than in stage IV. It also disagrees with CaPA over VI, where it shows lighter precipitation along the coastline and over southern portions of the island.

KED is similar to OK except that it adds a radar-based constraint to the weighting process: the radar-derived precipitation at each grid point must equal the weighted average of the radar values interpolated to the gauge locations. Areas of terrain blocking are omitted from the merging process, rendering the precipitation in these regions equivalent to the OK result. The resulting KED analysis for 14 February is nearly identical to the OK analysis (Fig. 4d), suggesting that the radar composite’s incomplete coverage over the PNW renders it unable to strongly constrain the result. We also compared the KED analyses with and without the snow undercatch correction described above. Again, the differences between the
two are small because only a small fraction of the non-SNOTEL gauges were subfreezing (not shown).

4. Verification metrics

a. Deterministic point verification

Although point verification can be misleading as a result of the double-penalty problem and the scale separation between a point location and model grid cell (e.g., Casati et al. 2008; Ebert 2009; Gilleland et al. 2009), past studies of orographic precipitation over western North America have used this approach (Colle et al. 2000; Mass et al. 2002; Grubišić et al. 2005). To permit comparison with those studies, we perform such an analysis herein. Additionally, because gridded precipitation estimates are highly uncertain in the PNW, point verification is attractive because it does not rely on such products. Figure 5 illustrates the ensemble-mean point precipitation forecasts (using nearest-neighbors

**Fig. 5.** The 24-h cumulative precipitation on 14 Feb 2011 for the (a) gauges and (b) HREnKF, (c) DS, and (d) REnKF ensemble means, with the latter three interpolated to the gauge locations using nearest-neighbor interpolation. Boxes are defined in Fig. 3.
interpolation) with the rain gauge observations on 14 February 2011, with the former restricted to the three predefined regions of highest gauge density. Because of the smoothing effects of ensemble averaging, the ensemble-mean fields exhibit lower intensities and less small-scale structure than individual ensemble members, particularly in the higher-resolution HREnKF and DS forecasts (e.g., Sercel et al. 2014).

The deterministic point verification employs the bias and equitable threat score (ETS), two widely used metrics (Schaefer 1990; Wilks 1995). Both are computed using a contingency table, where the sum of all hits \( A \), false alarms \( B \), misses \( C \), and correct negatives \( D \) is determined for each forecast based on a chosen precipitation threshold. The bias is the ratio of the frequency of predicted to observed events

\[
\text{Bias} = \frac{A + B}{A + C}, \tag{1}
\]

where values greater than 1, less than 1, and 1 indicate overprediction, underprediction, and perfect skill, respectively. The ETS, in contrast, quantifies the improvement of a forecast over a random forecast, the latter of which is quantified by the number of chance hits \( R \):

\[
R = \frac{(A + B)(A + C)}{A + B + C + D} \quad \text{and} \quad \tag{2}
\]

\[
\text{ETS} = \frac{A - R}{A - R + B + C}. \tag{3}
\]

An ETS of 1 indicates a perfect forecast, while an ETS of zero indicates no improvement over a random forecast. Although four thresholds were considered for both the bias and the ETS verification (1, 5, 10, and 20 mm), only the 10-mm results are shown herein for brevity, as the findings were consistent across all thresholds. Verification is conducted for all three ensembles and all seven events, over the three regions of interest (VI, PS, and VA). Total bias and ETS are also computed for the entire set of events using a total contingency table (Hamill 1999). The statistical significance of the differences in these total scores between different ensembles is evaluated using the Wilcoxon test, which determines if two random variables are drawn from continuous distributions with the same medians (Hamill 1999).

In both the bias \( 5 \) and ETS calculations, we treat each observation as independent and thus neglect the spatial correlation of errors. Because such correlations would tend to reduce our effective sample size, neglecting them may cause our statistical significance analyses to be overconfident. However, the small number of events sampled in this study hinders a rigorous analysis of these error correlations. We thus focus our discussion on differences in the distributions of the skill scores rather than the statistical significance of these differences.

A final component of our deterministic verification consists of Taylor diagrams (Taylor 2001), which display three different aspects of forecast skill in a single plot: 1) centered root-mean-square (RMS) difference between, 2) correlation between, and 3) standard deviation of observations and forecasts. These three quantities are plotted together on a quarter-circular diagram with standard deviation along the radius, correlation coefficient along the azimuth, and RMS error given by the distance between the forecast and the observation (the latter lying along the \( x \) axis).

b. Spatial verification

As discussed previously, spatial verification methods can be used to overcome the double-penalty problem (Gilleland et al. 2009). One simple such method is the fractions skill score (FSS; Roberts and Lean 2008), a “neighborhood” technique that evaluates forecast skill as a function of spatial scale. Three gridded analyses are chosen for spatial verification: stage IV, CaPA, and KED. Before computing the FSS, these products are first remapped onto the same grid as the lowest-resolution model or verification product being compared. For example, when comparing the HREnKF with the RenKF, the observations and HREnKF forecast are first smoothed to an effective resolution of 15 km and then interpolated onto the 15-km RenKF grid using a nearest-neighbor technique. Because the stage IV analysis is limited to the United States, we restrict all verification to the United States when this product is used. When the stage IV analysis is omitted and the KED and/or CaPA analyses are used, the full PNW may be considered.

The FSS is computed for a given precipitation threshold, which can be either a fixed accumulation or a percentile. Following Roberts and Lean (2008), we set the threshold for each case as the 90th percentile of the precipitating grid points. Then, the observational \( O \) and model \( M \) precipitation fields are converted into binary fields \( (I_o, I_m, \text{ respectively}) \) equal to 1 where precipitation exceeds the threshold and 0 elsewhere. Then, fields of fractions for both the observations and the models are created over a range of neighborhood grid lengths (defined by an integer number of grid points \( b \)). For each grid point, the fractions of points with \( I_o = 1 \) and \( I_m = 1 \) within the square neighborhood of \( b \times b \) grid points centered on that point are computed. The resulting fields of observed and model fractions are represented by \( O_{(b)}(i, j) \) and \( M_{(b)}(i, j) \), respectively, where \( i \) and \( j \) range from 1 to the domain size in \( x \) and \( y \) (\( N_x \) and

Unauthenticated | Downloaded 10/04/22 05:39 AM UTC
Fractions are computed at different scales by increasing $b$ up to $N_{\text{max}} - 1$, where $N_{\text{max}}$ is the longest horizontal dimension of the grid.

The FSS is defined as

$$\text{FSS}_{(b)} = 1 - \frac{\text{MSE}_{(b)}}{\text{MSE}_{(b)\text{ref}}},$$

where

$$\text{MSE}_{(b)} = \frac{1}{N_x N_y} \sum_{i=1}^{b} \sum_{j=1}^{b} \left[ O_{(b)(i,j)} - M_{(b)(i,j)} \right]^2$$

and

$$\text{MSE}_{(b)\text{ref}} = \frac{1}{N_x N_y} \left[ \sum_{i=1}^{b} \sum_{j=1}^{b} \left( O_{(b)(i,j)}^2 + \sum_{i=1}^{b} \sum_{j=1}^{b} M_{(b)(i,j)}^2 \right) \right],$$

and ranges from 0 to 1, with 0 the worst possible score and 1 the best. It tends to monotonically increase with $b$ as more small-scale precipitation features are admitted into the averaging. The model skill relative to a random forecast is obtained by comparing the FSS with $\text{FSS}_{\text{uniform}} = 0.5(1 + f_o)$ (where $f_o$ is the fraction of the domain where precipitation exceeds the threshold), with the latter representing the FSS that would be obtained at the grid scale by a random forecast with the same $f_o$. The scale at which the FSS crosses $\text{FSS}_{\text{uniform}} (\Delta_{\text{min}})$ is the minimum scale at which the forecast becomes more skillful than a random forecast.

Although the FSS was formulated for deterministic forecasts, it can be applied to an ensemble by computing it separately for each member, then analyzing the distribution of scores within each ensemble. Moreover, just as the FSS can be computed between models and observations to assess model skill, it can also be computed between ensemble members to assess ensemble spread (or “FSS spread”; Dey et al. 2014). In a similar fashion, we also use the FSS to quantify the uncertainty of the verification products themselves.

### Probabilistic verification

Probabilistic verification is conducted in the traditional way: at each grid point, the probability of detecting precipitation over a given threshold is defined as the proportion of all ensemble members forecasting rain exceeding that threshold at that point. Ensemble forecasts of daily precipitation accumulations are evaluated using three metrics: the rank histogram, the reliability diagram, and the relative operating characteristics (ROC) curve. These metrics offer complementary information on two independent qualities of ensemble systems: reliability and resolution (e.g., Talagrand et al. 1997).

A probabilistic forecasting system is reliable if the observed frequencies of occurrence of a given event (here, rain exceeding a certain threshold) are consistent with the forecast probabilities. If the system is reliable, an event forecast with a 30% probability should occur 30% of the time. One way of quantifying ensemble reliability is through rank histograms, which, if properly interpreted, are also useful for diagnosing errors in the mean and the spread of an ensemble prediction system (Hamill 2001). Rank histograms are obtained by counting, over many samples, the rank of an observation with respect to the corresponding forecast values from the ensemble sorted from low to high. For an ensemble of $N$ members, there are $N + 1$ ranks, and the frequency of each rank is then illustrated in a histogram plot. The shape of the resulting histogram gives information on the ensemble quality: a flat histogram indicates good ensemble reliability, a U-shaped histogram indicates an underdispersive ensemble as the observation falls at or near the extremes of the ensemble range, and a sloped histogram may indicate a consistent bias.

We produce rank histograms of 24-h precipitation accumulations for each ensemble system, using gauge data for verification for all seven events. Grid points where the verification equals some of the ensemble members, such as locations with no precipitation, are treated similarly to those of Hamill and Colucci (1997). For these points, the number of ensemble members $M$ with values equal to the verification is determined. Then, very small uniformly distributed random numbers (between 0 and 0.0004 mm) are added to the $M$ members and to the verification, and the rank is redetermined based on the resulting values. This procedure assigns a random rank to the observation, avoiding any biases created by always placing such observations in the same bin.

Another way of assessing ensemble reliability is through the reliability diagram, which plots the observed frequency of occurrence versus probabilities of some categorical forecast, such as the occurrence of precipitation accumulation exceeding a certain threshold. A reliable ensemble would produce points aligned along the diagonal. Since the observed frequency of occurrence of an event is computed for different probability bins (e.g., between 10% and 20%), it is important to consider the total number of samples in each bin. The number of data points available for a certain probability category is therefore illustrated by a complementary sharpness histogram. As mentioned by Wilks (1995), reliability diagrams must be carefully interpreted and can be misleading if computed over small sample sizes. For sufficiently large datasets, the reliability of probabilistic forecasts can be improved through postprocessing.

The reliability diagrams and sharpness histograms are computed for 24-h precipitation accumulations, verified...
against gauge observations at the gauge locations for the seven events. For each precipitation threshold, a probabilistic ensemble forecast is first obtained. Then, for each probability category $p_0 = 0.1, 0.2, \ldots, 1$, we find the data points $X_i$ where the probabilistic forecast shows fractional occurrences of $p_0 - 0.05 \leq p < p_0 + 0.05$. The number of such points is then recorded in the sharpness histogram. The observed frequency corresponding to this probability category is the proportion of these points where the observed accumulations exceed the threshold.

Similar to the reliability diagram, the ROC curve is computed for categorical precipitation forecasts to measure ensemble resolution. Ensemble resolution is related to the variability in the predicted probability distributions, or the ability of the ensemble to differentiate between more likely and less likely events. A simple method for quantifying ensemble resolution consists of plotting the ROC curve for different events (where an “event” is defined as the occurrence of precipitation above a certain threshold). For such an event, a probabilistic forecast is obtained from the ensemble and verified as follows: for a given probability threshold $p_0$, the number of hits, misses, false alarms, and correct negatives are counted and used to calculate the false alarm rate (FAR) and the probability of detection (POD). A hit is defined as a data point where the observed precipitation exceeds the threshold and $p \geq p_0$. A miss is associated with observed precipitation exceeding the threshold and $p < p_0$, a false alarm with observed precipitation below the threshold and $p \geq p_0$, and a correct negative with observed precipitation below the threshold and $p < p_0$. Then, FAR and POD are calculated as [using the same notation as in (1)–(3)]
POD is then plotted versus FAR for different \( p_0 \) values, with larger areas under the resulting curve indicating better ensemble resolution. If the resulting points are located along the diagonal, the ensemble has no resolution. Again, 10 probability values \( p_0 \) = 0.1, 0.2, ..., 1 are used to verify 24-h precipitation accumulations forecasts from the three ensembles against gauge observations for all seven events at different precipitation thresholds.

5. Verification results

As a first step in the verification, we qualitatively compare ensemble-mean 24-h precipitation forecasts with observational analyses for the four heaviest precipitation events (4, 12, 13, and 14 February 2011; Figs. 6–9). The analyses generally indicate precipitation enhancements over the Olympic and Cascade ranges in most events. However, they substantially differ from each other in some cases, with the CaPA appearing as a clear outlier on 12 and 13 February. (Figs. 7 and 8). Differences between the stage IV and KED are generally smaller but still noticeable. Of the three observational products, the forecasts tend to best match the stage IV analyses over the U.S. portion of the grid, which is encouraging given our higher level of confidence in that product. Compared with the HREnKF and DS forecasts, which are nearly indistinguishable in most cases, the REnKF forecasts clearly contain smoother and coarser precipitation fields, and slightly more domain-averaged precipitation.
a. Deterministic verification results

The bias scores at the 10-mm threshold fluctuate greatly between the seven events, which overwhelm any systematic differences between the ensembles (Fig. 10). A clearer picture is provided by the total bias in the rightmost column, where the REnKF exhibits positive bias over PS and VA. Over these regions, the REnKF bias is larger than those of the HREnKF and DS ensembles, which lie much closer to unity. Such errors may stem from a number of factors including the smoother terrain field of the REnKF and its diagnostic microphysics scheme, which neglects hydrometeor advection (e.g., Lean and Browning 2013). In contrast, the REnKF ensemble slightly outperforms both the HREnKF and DS over VI, where all three ensembles exhibit a weak negative bias. This improvement, however, is small compared with the REnKF skill degradations over the VA and PS regions. Relative to the HREnKF, the DS also exhibits a slight bias improvement over VI, as well as a slight degradation over VA. Note that the PS region only experiences substantial rainfall on 13 and 14 February; hence, the results there should be considered with caution.

Compared with the bias, the ETS at the 10-mm threshold indicates a clearer improvement of the HREnKF and DS over the REnKF ensemble, not only in the total ETS but also in most individual events, over all three regions (Fig. 11). Also, more spread is evident in the skill of the high-resolution ensemble members than for the REnKF ensemble members, a point that will be revisited in section 5c. While the differences between the HREnKF and REnKF are statistically significant over all three regions, those between the HREnKF and DS are insignificant. Taken together, the bias and ETS suggest a robust improvement in forecast...
skill owing to the higher grid resolution of the HREnKF and DS systems, but with no obvious improvement stemming from the high-resolution data assimilation in the HREnKF.

Taylor diagrams for the four heaviest precipitation events in Fig. 12 indicate clusters of densely packed points for each forecasting system during each event, reflecting a lack of ensemble dispersion. Nonetheless, the forecast skill, as measured by the distance between the forecast points on the diagram and the observation point, is generally higher for the HREnKF and DS forecasts than for the REnKF. The distributions of points within the HREnKF and DS ensembles are similar enough to have been drawn from the same probability density function. The only event where the REnKF outperforms the higher-resolution ensembles in any forecast aspect is 13 February (Fig. 12c), where the forecast standard deviation more closely matches that of the observations.

We first use the FSS to assess the differences between our three gridded precipitation analyses. Figure 13 shows the FSS for stage IV compared with KED (stage IV-KED), CaPA (stage IV-CaPA), and the HREnKF ensemble mean (stage IV-HREnKF) for the four heaviest precipitation days. The stage IV-HREnKF result provides a useful reference point for establishing whether the analyses agree better with each other than with the forecasts. On 12 February, the FSS of stage IV-HREnKF exceeds that of stage IV-CaPA and stage IV-KED (Fig. 13b), with $\Delta_{\text{min}} = 2.5$ km (the grid scale) compared with 25 km (stage IV-KED) and 225 km (stage IV-CaPA). These differences likely stem from the lighter precipitation over the western Olympics in the KED and CaPA products. During the other three events, the forecast skill (as quantified by the stage IV-HREnKF FSS) is lower than the difference between stage IV and
KED (as quantified by the stage IV-KED FSS). Given the generally poor agreement between CaPA and stage IV, we henceforth omit the CaPA from the FSS analysis.

The FSS for the three ensembles, verified using the stage IV analysis, are compared in Fig. 14. Recall that, because the stage IV product is used and REnKF forecasts are included, the verification grid is restricted to the U.S. subdomain and uses a spacing of 15 km. Overall, the HREnKF and the DS ensembles perform very similarly on this grid and outperform the REnKF on 4,

FIG. 10. Bias in 24-h cumulative precipitation for the HREnKF (blue), DS (red), and REnKF (orange) results computed against rain gauges over the (a) VI, (b) PS, and (c) VA regions, at a 10-mm threshold. The total bias is shown in the far-right column, where the boldface box plots represent a statistically different result than those of HREnKF. The box plots indicate the ensemble median (central line), 25th and 75th percentiles (bottom and top box edges), and maximum and minimum (whiskers). The values of \( n \) and \( p \) provided for each event, respectively, correspond to the number of gauge observations in that region and the number of those gauges with precipitation exceeding the chosen threshold.

FIG. 11. As in Fig. 10, but for the ETS verification.
12, and 14 February. On these days, $\Delta_{\text{min}}$ is near the grid scale (2.5 km) for both the HREnKF and the DS but much higher (50 km) for the REnKF.

The only day in which the REnKF outperforms any other ensemble is 13 February (Fig. 14c), where it scores higher than the DS but not the HREnKF. The generally poor scores on this day stem from the meteorological conditions: rather than being stratiform as on the other three days, the precipitation consisted of shallow cellular convection in postfrontal flow (not shown), which is generally more chaotic and less regulated by the terrain. The relatively large FSS discrepancy between the DS and HREnKF in this case is attributable to the cold start of the DS ensemble and the fact that most precipitation occurred over 0000–0600 UTC (not shown). The DS produced nearly zero precipitation over the first hour, which caused it to underpredict the precipitation, most noticeably over and to the south of the Olympics (Fig. 8). Although the REnKF also uses a cold start and underpredicts precipitation over the first hour, its tendency to overpredict precipitation offsets this effect to give a more accurate forecast than the DS ensemble.

If the observations were sampled from the same probability distribution function as the ensemble members, the forecast error would equal the spread of the
members, when averaged over many cases. The level to which the individual ensemble forecasts achieve this desirable “spread–skill” relation may be qualitatively evaluated by comparing the FSS computed for pairs of HREnKF ensemble members (HREnKF spread, which is similar to or larger than the spreads of the DS and REnKF ensembles) to that between the different ensembles and the observations. On 4 and 13 February, the ensemble errors greatly exceed the HREnKF spread, indicating strong underdispersion (Fig. 14c). This finding is consistent with previous verification studies of convective-scale ensembles (e.g., Vié et al. 2011; Surcel et al. 2015).

By contrast, on 12 and 14 February the ensembles are all very accurate, and the forecast error is similar to the HREnKF spread at scales above 100 km.

To summarize the FSS results over the seven events, we present FSS aggregates in Fig. 15a. For percentile thresholds like that used here, the aggregate FSS is simply the average FSS across the events (Roberts 2008). Consistent with the individual event results in Fig. 14, the HREnKF exhibits the highest aggregate FSS, followed by the DS, and then the REnKF. The difference between the HREnKF and DS is owing partly to the former’s data assimilation and partly to the latter’s aforementioned spinup issues, both of which are most prevalent early in the forecasts. This point is clearly seen from the FSS aggregates over two different subintervals: 0000–0600 and 0600–0000 UTC (Figs. 15c,d). Because the aggregates weight each event equally, care must be taken to omit subperiods with minimal precipitation. Thus, we restrict the analysis to events where the maximum stage-IV-observed precipitation exceeded 5 mm (15 mm) over 0000–0600 UTC (0600–0000 UTC), which filters out two events from each subperiod. While the marginal improvement of the HREnKF over the DS is mainly realized over the first 6 h of the forecast, the positive impacts of higher resolution are evident over all time periods.

The FSS over the entire PNW domain, evaluated using the KED analysis (Fig. 16b), is broadly consistent with the stage IV analysis except that the scores are generally worse (and $\Delta_{\min}$ is larger). This reduction in FSS is not necessarily associated with increased forecast error but with the aforementioned uncertainties of the

---

**Fig. 13.** FSS for the stage IV vs the KED (stage IV-KED), CaPA (stage IV-CaPA), and HREnKF 24-h forecasts (stage IV-HREnKF) for the 90th percentile of 24-h precipitation on (a) 4, (b) 12, (c) 13, and (d) 14 Feb 2011. The stage IV-HREnKF curve represents the mean FSS for the ensemble, along with error bars of one standard deviation.
KED verification product, particularly over VI, where it disagrees strongly with both the CaPA and the model forecasts (Figs. 6–9). By contrast, if the KED verification is restricted to the United States (i.e., the stage IV grid), the scores improve slightly, to a level more consistent with the stage IV verification (cf. Figs. 16a and 14). Over both regions, the HREnKF and DS again perform very similarly and outperform the REnKF.

FIG. 14. FSS for the HREnKF, DS, and REnKF ensembles, along with HREnKF spread, on (a) 4, (b) 12, (c) 13, and (d) 14 Feb 2011. The FSS was computed using the stage IV verification product on the stage IV (U.S.) grid, smoothed to the REnKF 15-km grid. The value of the 90th percentile of precipitation on that day is \( P \). The curves represent the mean FSS for the ensemble, along with error bars of one standard deviation.

FIG. 15. Aggregate FSS analyses over all seven events using the stage IV verification product over (a) 0000–0000, (b) 0000–0600, and (c) 0600–0000 UTC. Thresholds of 5 mm (15 mm) were applied to omit very light precipitating events in (b) and (c).
To determine whether higher grid resolution improves the forecasts at small scales, we also verified the HREnKF and DS against the stage IV product at the models’ 2.5-km native grid spacing (rather than the coarser REnKF grid). This analysis reveals a modest improvement in FSS for the most poorly forecast event (13 February) over that on the coarser grid (not shown). Specifically, the HREnKF and DS reach FSS\textsubscript{uniform} at slightly smaller scales (15 and 70 km) than before (40 and 110 km). This improvement may stem from an improved representation of small-scale processes (e.g., lee waves and cumulus convection) that was previously masked by smoothing to a coarser grid.

c. Probabilistic verification results

Two main findings can be drawn from the results presented thus far: 1) the higher-resolution HREnKF and DS provide more accurate 24-h precipitation accumulation forecasts than the lower-resolution REnKF and 2) the spread among the ensemble members is smaller than the forecast error, indicating a general lack of ensemble reliability. More detail about this latter finding is provided by the probabilistic evaluation of 24-h precipitation accumulation forecasts, which is carried out as a point verification against precipitation gauges. All the gauges available for all seven events are used in the evaluation and are considered to be independent samples for a total of 2500 data points. The forecast values at rain gauge locations are obtained using nearest-neighbor interpolation. No spatial averaging is performed on the forecast precipitation fields prior to remapping. Because of the limited size of the dataset, the geographical dependence of the ensemble performance is not addressed here as in section 5a.

Figure 17 shows rank histograms for the three ensembles, all of which exhibit U shapes indicative of strong underdispersion consistent with the Taylor diagrams and FSS analysis. To quantify the degree of underdispersion, we compute $f_1$ and $f_{41}$, the proportion of data points where the observations, respectively, fall below the first and above the 40th sorted ensemble members. For all ensembles, $f_1 + f_{41}$ exceeds 0.5, indicating that over half of the observations fall outside the ensemble spread [similar underdispersion was seen in the HREnKF surface-station verification of Jacques et al. (2017)]. This sum is similar for the HREnKF and DS (0.55 and 0.56) and larger for REnKF (0.66), indicating slightly increased reliability at higher grid resolution.

Figure 18a shows the reliability diagram for all three ensembles, again for the 10-mm threshold. While the observed frequency of occurrence increases with ensemble probability for all ensemble systems, the two are not equal and thus the reliability is imperfect. From the sharpness histogram (inset in Fig. 18a), it is clear that very few data points (less than 0.04%) have probabilities between 0.2 and 0.8, which complicates the interpretation of the reliability diagram for these probability categories. Most data points exhibit a probability of either 0 or of 1, which is inconsistent with the observed frequency of occurrence and again reflective of ensemble underdispersion. The marginal improvement in performance in the HREnKF and DS is further reinforced by these diagrams. Analysis at other
precipitation thresholds (1 and 5 mm) indicates consistent results (not shown).

The issue of ensemble resolution is addressed through ROC curves, shown for the 10-mm threshold in Fig. 18b. The area under the curve exceeds 0.5 for all ensembles, indicating some ensemble resolution. The REnKF shows a clearly poorer performance than the HREnKF and DS, while the latter two ensembles are practically indistinguishable. For a threshold of 1 mm (essentially a rain/no-rain threshold for this accumulation length), the three systems have identical ROC curves (not shown).

6. Discussion

Two objectives of this study were to thoroughly verify 24-h HREnKF precipitation forecasts and to quantify any improvements of the HREnKF forecasts over corresponding forecasts from less computationally expensive alternatives (DS and REnKF). The central finding is that the 2.5-km HREnKF outperforms the 15-km REnKF, for nearly all verification metrics. Deterministic point verification using precipitation gauges revealed that the total bias and total ETS (the cumulative bias and ETS over all seven events) at the 10-mm threshold were all improved in the HREnKF. The lone exception to that rule was the bias score over Vancouver Island, where all three ensembles performed well. Similarly, Taylor diagrams indicated improvements in multiple verification aspects for the HREnKF and DS, relative to the REnKF. These improvements in point verification scores as the horizontal grid spacing is reduced from $O(10)$ km to $O(1)$ km are consistent with previous NWP studies over mountainous terrain (e.g., Colle et al. 2000; Mass et al. 2002; Schwitalla et al. 2008).

Because deterministic point verification is prone to misinterpretation, we also performed spatial verification using the FSS method. The FSS was used not only to assess the scale dependence of the forecast skill but also to gain insight into the ensemble reliability. Again, uniformly higher skill was obtained for the HREnKF than for the REnKF, regardless of the gridded observational analysis or verification grid used. The HREnKF achieved higher FSS scores at nearly all scales and became skillful at smaller scales than the REnKF. These absolute FSS scores were larger than those reported from the instantaneous verification of summer convection events in the United Kingdom (Roberts and Lean 2008), likely because of the more stratiform nature of the PNW winter precipitation, the more predictable forcing of its rugged terrain, and the longer 24-h accumulation window. While the FSS analysis suggested that all three ensembles were underdispersed, the HREnKF apparently suffered less from this problem than the REnKF.

Further insight into ensemble reliability (or dispersion) was provided by probabilistic verification using the rank histogram and reliability diagrams, which reinforced the underdispersive nature of the ensembles. Again, this analysis found the HREnKF underdispersion to be less severe than that of the REnKF. However, as mentioned before, each forecast was evaluated at its native resolution. Therefore, it is not possible to determine whether the improved verification results of the higher-resolution models are achieved as a result of a more correct representation of precipitation physics, or because of the high-resolution models simply
having a more realistic representation of the probability distribution function of the precipitation amounts. In addition, while it is clear that higher-resolution ensembles are more suitable for providing probabilistic forecasts of precipitation in the PNW, more work is necessary to improve the probabilistic forecasts in this region. Because the ensembles are currently not reliable, this could be addressed by calibrating the ensemble forecasts through postprocessing (Hamill et al. 2008). However, for successful ensemble calibration, both a sufficiently large dataset and high quality verification data are necessary, which represented two major shortcomings of the present project.

In general, the DS and HREnKF ensembles performed similarly. Although the HREnKF exhibited slightly higher verification scores, most of these differences were not statistically significant. Moreover, the small discrepancies between the DS and the HREnKF stemmed in part from the former’s cold start, where all hydrometeor species were reset to zero at initialization and regenerated during model spinup. Because this issue was inherited from the REnKF, both the REnKF and DS forecasts suffered from it. In the REnKF, however, this problem was partially offset by a tendency to overpredict precipitation. We speculate that the inability of the HREnKF forecasts to significantly improve upon the DS stems from the lack of upstream observations over the Pacific Ocean and the sparsity of the observations that were assimilated over the PNW (Jacques et al. 2017). Future work will address objectives 2 and 3 mentioned in section 1 by assimilating high-resolution reflectivity and Doppler velocity data from regional radars and more closely analyzing the time evolution of forecast skill.

Another objective was to assess the uncertainty of the available gridded observational products over the PNW, which are challenged by the sparse rain gauge observations and widespread radar beam blocking in this region (e.g., Westrick et al. 1999). We compared three such products, the U.S. stage IV analysis, the Canadian Precipitation Analysis (CaPA), and our own region-wide analysis using kriging with external drift. The differences between these products were substantial, occasionally exceeding those between a given product and the model forecasts, which adds uncertainty to the FSS analyses.

7. Conclusions

This study has verified 24-h precipitation forecasts from three experimental ECCC ensemble forecast systems: the 2.5-km high-resolution ensemble Kalman filter (HREnKF), the 15-km regional ensemble Kalman filter (REnKF), and a third ensemble downscaled from the REnKF with the same high-resolution grid as the HREnKF (DS). Forecasts for seven heavy orographic precipitation events over the first half of February 2011 were conducted with ECCC’s GEM model. The objectives of this work were to thoroughly evaluate the ensemble forecast skill, quantify any marginal benefits
of the HREnKF over less expensive alternatives, and estimate the uncertainty of available gridded observational products needed for spatial verification.

Quantitative precipitation forecasts (QPFs) from the HREnKF, DS, and RENKF systems were evaluated against precipitation observations using a range of verification methods. Deterministic verification using the bias, equitable threat score (ETS), Taylor diagrams, and fractions skill score (FSS) all showed the HREnKF and DS models performing similarly to each other and generally outperforming the RENKF, both in terms of absolute skill and ensemble reliability. Probabilistic verification results obtained from rank histograms, reliability diagrams, and ROC plots indicated that, while all ensembles suffered from underdispersion, this problem was more severe for the RENKF. Although more accurate probabilistic forecasts over this region could be achieved through the postprocessing of precipitation forecasts from the high-resolution ensembles, such an analysis was not attempted here.

Most of the benefits achieved by the HREnKF over the REnKF were also realized by the DS system, which does not use the very costly continuously cycled convective-scale data assimilation that limited the overall length of the analysis period to half of February 2011. Thus, the DS approach appears to be superior to the HREnKF in terms of performance for cost. However, high-resolution data assimilation over the PNW may still be worthwhile if it uses the full suite of observations available, in particular operational radar data. Since late 2011, a U.S. NEXRAD radar has been operational at Langley Hill, Washington, providing a view of southwestern Washington that herein was obscured by the Olympic Mountains.

The FSS verification relied on three gridded observational products that all suffer from major limitations over the PNW. Although the more advanced radar processing and radar–gauge merging of the U.S. stage IV product likely renders it the most accurate of the three, it remains uncertain over complex terrain, and its geographic limitation to the United States diminished its utility. While the CaPA covers the full PNW, it disagreed more strongly with the forecasts themselves. Thus, we created our own PNW-wide analysis using kriging with external drift to merge radar and rain gauge data, which agreed better with the stage IV analyses. However, this product appeared to be unreliable over Vancouver Island, likely because of the scarcity of gauges there. Such findings emphasize the need for a reliable gridded precipitation product covering the whole PNW.

**Acknowledgments.** Funding for this research was provided by the Marine Environmental Observation Prediction and Response (MEOPAR), through Grant ECI-DK-MCG. The probabilistic evaluation methodology was developed by MS during a postdoctoral appointment financially supported by Prof. M. K. Yau through a Natural Science and Engineering Research Council (NSERC) Industrial Research Chair (IRC) grant.

**REFERENCES**


