The utility of convection-permitting ensembles for the prediction of stationary convective bands

The Utility of Convection-Permitting Ensembles for the Prediction of Stationary Convective Bands

ANDREW I. BARRETT AND SUZANNE L. GRAY

Department of Meteorology, University of Reading, Reading, United Kingdom

DANIEL J. KIRSHBAUM

Department of Atmospheric and Oceanic Sciences, McGill University, Montreal, Quebec, Canada

NIGEL M. ROBERTS

MetOffice@Reading, University of Reading, Reading, United Kingdom

DAVID M. SCHULTZ AND JONATHAN G. FAIRMAN JR.

Centre for Atmospheric Science, School for Earth, Atmospheric and Environmental Sciences, University of Manchester, Manchester, United Kingdom

(Manuscript received 7 April 2015, in final form 15 October 2015)

ABSTRACT

This study examines convection-permitting numerical simulations of four cases of terrain-locked quasi-stationary convective bands over the United Kingdom. For each case, a 2.2-km-grid-length, 12-member ensemble and a 1.5-km-grid-length deterministic forecast are analyzed, each with two different initialization times. Object-based verification is applied to determine whether the simulations capture the structure, location, timing, intensity, and duration of the observed precipitation. These verification diagnostics reveal that the forecast skill varies greatly between the four cases. Although the deterministic and ensemble simulations captured some aspects of the precipitation correctly in each case, they never simultaneously captured all of them satisfactorily. In general, the models predicted banded precipitation accumulations at approximately the correct time and location, but the precipitating structures were more cellular and less persistent than the coherent quasi-stationary bands that were observed. Ensemble simulations from the two different initialization times were not significantly different, which suggests a potential benefit of time-lagging subsequent ensembles to increase ensemble size. The predictive skill of the upstream larger-scale flow conditions and the simulated precipitation on the convection-permitting grids were strongly correlated, which suggests that more accurate forecasts from the parent ensemble should improve the performance of the convection-permitting ensemble nested within it.

1. Introduction

Quasi-stationary convective bands regularly develop over and/or downwind of complex topography, where stationary updrafts are generated by prominent orographic features and/or land surface variations. Observational evidence suggests that these bands may produce heavy localized precipitation and, in some cases, flash flooding. Two recent examples of catastrophic flooding from such bands in the United Kingdom include the Boscastle flood of August 2004 (Golding 2005) and the Ottery St. Mary hailstorm in October 2008 (Clark 2011). Other heavy precipitation events associated with terrain-locked bands have been reported over Japan (Yoshizaki et al. 2000), southern France (Miniscloux et al. 2001; Cosma et al. 2002), and the U.S. Pacific Northwest (Kirshbaum and Durran 2005). The physical mechanisms anchoring the bands include, among others, gravity waves

Denotes Open Access content.

Corresponding author address: Andrew Barrett, Dept. of Meteorology, University of Reading, P.O. Box 243, Earley Gate, Reading RG6 6BB, United Kingdom.
E-mail: a.i.barrett@reading.ac.uk

DOI: 10.1175/MWR-D-15-0148.1

and/or lee-side convergence past local terrain ridges (e.g., Mass 1981; Cosma et al. 2002; Kirshbaum et al. 2007a; Barrett et al. 2015) and quasi-stationary sea-breeze fronts (e.g., Warren et al. 2014). In such events, the complexity of the underlying terrain may exacerbate the flash-flooding risks by channeling precipitation through steep-sided valleys into narrow water catchments.

Because of their potential for heavy precipitation, quasi-stationary convective bands represent an important forecasting problem. Until recently, however, the narrowness of these bands [~2–10 km, according to Kirshbaum et al. (2007b)] rendered them unresolvable in regional forecast models with grid spacings of $O(10)$ km. Only in the past 10 years have $O(1)$-km convection-permitting grids emerged that offer the hope of explicitly capturing the bands (e.g., Lean et al. 2008; Stensrud et al. 2009). However, given the narrowness of the bands, they remain poorly resolved even on $O(1)$-km grids.

Since their origin, convection-permitting models have been mainly used "deterministically" to provide single forecast realizations. Although such forecasts provide valuable finescale detail and eliminate the need for a deep-convection parameterization scheme, they do not account for initial-condition uncertainties and model errors that cause forecasts to diverge from reality. Thus, attention is increasingly being focused on convection-permitting ensembles, which incur a much larger cost but attractively provide explicit and probabilistic storm prediction simultaneously. Various experiments with such ensembles have highlighted their advantages over deterministic or coarser-resolution ensemble forecasts for predicting convective precipitation (e.g., Kong et al. 2006, 2007; Clark et al. 2009, 2010, 2012). For example, in case studies of two convection events over central Europe, Hanley et al. (2011, 2013) found that some ensemble members provided far more accurate predictions of the event than did the "control" members, which would have served as the sole realizations of deterministic forecasts. Similarly, simulations of nonorographically forced snowbands have highlighted that changing the initial conditions, and hence the large-scale environment, can alter the organization of the simulated precipitation (Suarez et al. 2012).

Although convection-permitting ensemble forecasts represent an exciting new forecasting technology, computational constraints still limit their potential usefulness. One of the greatest benefits of these ensembles is their capability to provide guidance on potentially high-impact precipitation events that may not be captured in deterministic forecasts. However, the large ensemble sizes that are required to capture low-predictability events, particularly those characterized by small spatial scales (Richardson 2001; Clark et al. 2011), are often unfeasible operationally. As an alternative, one may artificially increase the ensemble size by including the members of an ensemble initialized a few hours earlier. However, if the statistics of the two ensembles are significantly different, their merger cannot be expected to accurately represent the distribution of possible outcomes.

Another constraint on the skill of a convection-permitting forecast is the skill of the parent forecast in which it is embedded. Although the skill of the parent forecasts is known to influence that of their nested ensembles (e.g., Roebber et al. 2008; Hanley et al. 2011; Novak and Colle 2012; Hanley et al. 2013), this relationship is not always straightforward. In a case study of a terrain-locked convective band downwind of the U.K. Lake District, Barrett et al. (2015) found convection-permitting forecasts of the band to be more skillful when the impinging flow upstream of the Lake District was represented more accurately. However, the skill was not well correlated with the winds over larger regions such as the whole United Kingdom or the whole model domain. In a study of Mediterranean high-precipitation events, Vié et al. (2011) found that model initial conditions had a strong impact in the first 12 h of their simulations, but the magnitude of the impact was strongly dependent on the synoptic situation.

Although convection-permitting ensembles have been verified in case studies and over broad samplings of convection events (e.g., Elmore et al. 2003; Vié et al. 2011), they have not been rigorously verified with respect to specific weather phenomena. This motivates the current study, which assesses the skill of these ensembles at representing one mode of potentially high-impact weather: terrain-locked convective bands. In particular, we study four recent such events in the United Kingdom to determine whether convection-permitting ensemble simulations succeed in accurately representing the bands. Specifically, we address the following questions:

1) Do convection-permitting ensembles capture the structure, location, timing, intensity, and duration of quasi-stationary convective bands?
2) What evaluation methods provide useful insights into forecast skill for these events?
3) Is there a strong correlation between the skill of the parent ensemble members and the convection-permitting ensemble members nested within them?
4) Can the ensemble size for these events be increased by using time-lagged ensembles?

To address these questions, 12-member convection-permitting (2.2-km grid spacing) ensemble simulations are analyzed for each event, using two different initialization times. For the sake of comparison, higher-resolution deterministic simulations (1.5-km grid spacing)

from the same two initialization times are also analyzed. The convection events, numerical model, and data sources are summarized in section 2. Section 3 outlines the object-based verification diagnostics used to evaluate the model simulations. Section 4 interprets the diagnostics in the context of one of the four cases; the remaining three cases are summarized in section 5. Section 6 addresses the utility of time-lagged ensembles and the relationship between skill in the convection-permitting ensemble and the larger-scale driving ensemble. Section 7 concludes the paper.

2. Overview of cases and data sources

a. Case descriptions

The four cases under investigation all occurred during the second half of 2012 and represent the most prominent quasi-stationary, terrain-locked convection events for which archived model data were available (July 2011–December 2012). The bands in all cases are assumed to be convective (either isolated or embedded) because of their strong similarities (in both structure and evolution) to previously observed terrain-locked convective bands (e.g., Miniscloux et al. 2001; Kirshbaum and Durran 2005). Although the horizontal scales of the bands were similar in all cases (5–10 km in width; 40–80 km in length), their location, intensity, duration, and stationarity varied from case to case. Although none of these events was particularly severe, they still represent useful cases for testing the model representation of terrain-locked convective bands.

Case 1, on 26 August 2012, featured a band anchored at its upstream end near the west coast of central England (Fig. 1b). This so-called "Cheshire Gap" event (Browning et al. 1985) consists of flow channeled over the Cheshire plain between two areas of elevated terrain—the Welsh mountains to the southwest and the Pennines to the northeast. In northwesterly flow, as is the case here (Fig. 1a), a convective band may initiate near the coastline and extend inland. The band persisted for 7 h, over which time the maximum radar-derived precipitation accumulation was 52 mm.

Case 2 occurred just two days after case 1, but over the Welsh mountains in southwesterly flow after a cold-frontal passage (Fig. 1c). Several flow-parallel bands of modest intensity formed over the windward slopes of the Welsh mountains (Fig. 1d), which strongly resemble those observed over the Oregon Coast Range in Kirshbaum and Durran (2005). Although the individual bands persisted for only 1–2 h, banded convection prevailed in the area for 4 h. The rain rates from these bands were typically only 2–6 mm h^{-1}, but their stationarity led to total accumulations of 12 mm.

Case 3, on 9 September 2012, involved a single flow-parallel band to the south of the Lake District in southwesterly flow ahead of an approaching Atlantic low-pressure system (Figs. 1e,f). This band was present for 3 h and was accompanied by lighter, nonstationary rain in the surrounding areas. The maximum precipitation accumulation associated with this band was 14 mm.

Case 4, on 29 December 2012, consisted of a quasi-stationary flow-parallel band over the Great Glen, a narrow valley that spans the width of Scotland (Fig. 1h). Other less intense bands were also aligned with, but located to the southeast of, the main band. The main band was present for 6 h, with a peak accumulation of 20 mm. The bands developed after a cold-frontal passage (Fig. 1g) and were ultimately disrupted by the approach of the occluded front from the northwest.

For each event, we define a band-centered verification box to focus our evaluation of the simulated precipitation. The verification box size (60 km wide and 220 km long) is the same for each case, centered on and aligned with the main band(s) (Fig. 2). The box size was chosen to be large enough to capture the observed banded precipitation for all of the events and to tolerate some misplacement of the precipitation by the model, but small enough to focus primarily on the bands of interest. Naturally, there is a trade-off between these two factors and the quantitative verification will be sensitive to the size of verification box chosen. Nonetheless, the chosen size yields quantitative results that are consistent with our qualitative characterization of the model performance. The period of model evaluation, which differs from case to case, spans from 2 h before the observed band(s) formed until 2 h after it dissipated.

b. Rainfall observations and forecast verification

The total precipitation accumulation was derived from the Met Office radar network rain-rate product, which is updated every 5 min on a fixed 1-km grid covering the United Kingdom. Rain rate is derived using radar reflectivity measured from the nearest radar using a calibration based on nearby rain gauges (Harrison et al. 2009, 2012). Processing removes ground clutter and other spurious returns and also corrects for seeder–feeder orographic enhancement beneath the radar beam (Harrison et al. 2009, 2012). This product offers the best estimate of precipitation distribution

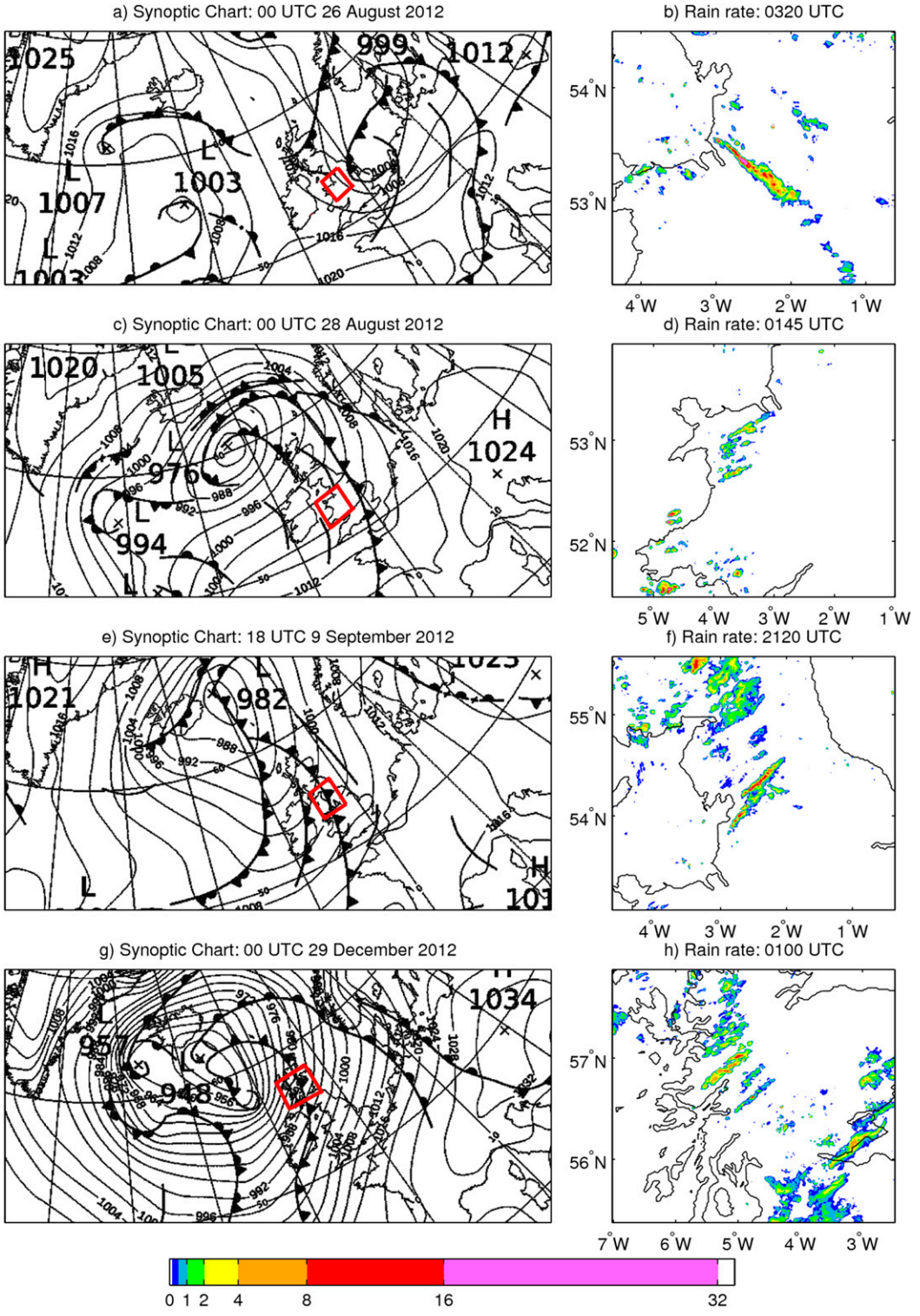

FIG. 1. Met Office synoptic analysis charts (Crown copyright, Met Office) from (a),(c),(e),(g) the nearest time to band formation and (b),(d),(f),(h) instantaneous radar-derived rain rate (mm h^{-1}) showing band structure. Each row is for a different case. The red boxes in the left panels mark the approximate area of the zoomed area in the right panels.

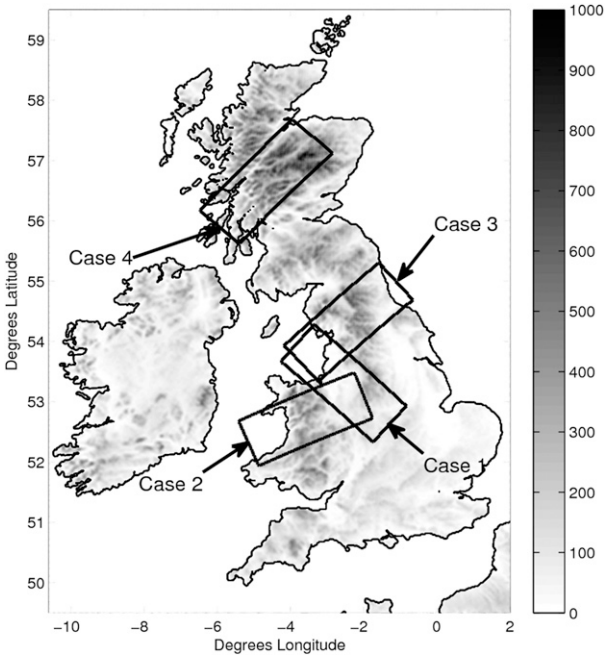

FIG. 2. Location of the analysis regions for the four cases. The terrain height (m) from the 2.2-km-grid-length model is shown by the shading, with the coastline marked in black.

across the United Kingdom and matches gauge accumulations well when averaged over several years (Fairman et al. 2015). For model evaluation, the radar data are mapped onto the model grid (1.5- or 2.2-km grid spacing) using a nearest-neighbor interpolation.

c. *Model description and setup*

All model simulations herein are operational simulations of the Met Office Unified Model (MetUM), which the Met Office runs on domains ranging from the entire globe to limited-area domains just encompassing the United Kingdom (Brown et al. 2012). The MetUM solves the nonhydrostatic, fully compressible deep-atmosphere equations of motion using semi-implicit, semi-Lagrangian time integration (Davies et al. 2005). Physical parameterizations include two-stream radiation (Edwards and Slingo 1996), subgrid cloud (Smith 1990), and mixed-phase microphysics (Wilson and Ballard 1999; including prognostic rain at convection-permitting resolutions). The Lock et al. (2000) boundary layer scheme is used for vertical mixing with a two-dimensional Smagorinsky (1963) mixing scheme in the horizontal. The Gregory and Rowntree (1990) convection scheme is used in the global and regional ensemble simulations, but not for convection-permitting simulations.

We use operational MetUM output from the Met Office Global and Regional Ensemble Prediction System (MOGREPS; Bowler et al. 2008), which, at the time of our cases, produced 12 ensemble members (a control and 11 perturbed members). A convection-permitting (2.2-km grid length) ensemble (MOGREPS-UK) is nested within the regional ensemble (MOGREPS-R; 18-km grid length); the domains of these simulations are shown in Fig. 3. The regional ensemble members are themselves nested within the global ensemble (MOGREPS-G; 60-km grid length). Each MOGREPS-UK simulation has 70 stretched vertical levels. The model lid is at 80 km for the MOGREPS-G and MOGREPS-R simulations but at 40 km for the MOGREPS-UK simulations.

MOGREPS-G simulations are initialized every 6 h, with the MOGREPS-R simulations initialized 6 h later and the MOGREPS-UK simulations initialized 3 h after that (Mylne 2013). Initial conditions for the control members of MOGREPS-G and MOGREPS-R are provided by four-dimensional variational data assimilation (4D-VAR). Perturbed members in MOGREPS-G are created by an ensemble transform Kalman filter [see Bowler et al. (2009) and Bowler and Mylne (2009) for details]. In MOGREPS-R, perturbations to the analysis are calculated as the difference between the perturbed and control members in MOGREPS-G at $T + 7$. These perturbations are applied to MOGREPS-R over a 2-h period equivalent to $T + 6$ to $T + 8$; hence, differences are calculated at $T + 7$. The initial conditions for MOGREPS-UK are downscaled (interpolated to a higher-resolution grid) from the corresponding MOGREPS-R ensemble member; no additional data assimilation is included. The MOGREPS-UK ensemble members take around 4–6 h to spin up features on the grid scale. By contrast, the deterministic UK variable resolution (UKV) model with 1.5-km grid length over the United Kingdom is initialized using 3D-VAR data assimilation (Tang et al. 2013), which includes nudging using radar rain rates. In these simulations, the sea surface temperatures are prescribed from a daily climatology and the soil moisture is an analyzed field.

For each case, we analyze convection-permitting forecasts (both deterministic and ensemble) that were initialized at two different times: approximately 12 h ($t - 12$) and 18 h ($t - 18$) before the band formed. The grid spacing of the convection-permitting forecasts (1.5 and 2.2 km) is likely insufficient to adequately represent bands that are often 5–10 km across. Nevertheless, our aim is to determine whether the current suite of convection-allowing operational models can predict these types of systems.

3. Verification diagnostics

Traditional measures of skill for model quantitative precipitation forecasts are not satisfactory for

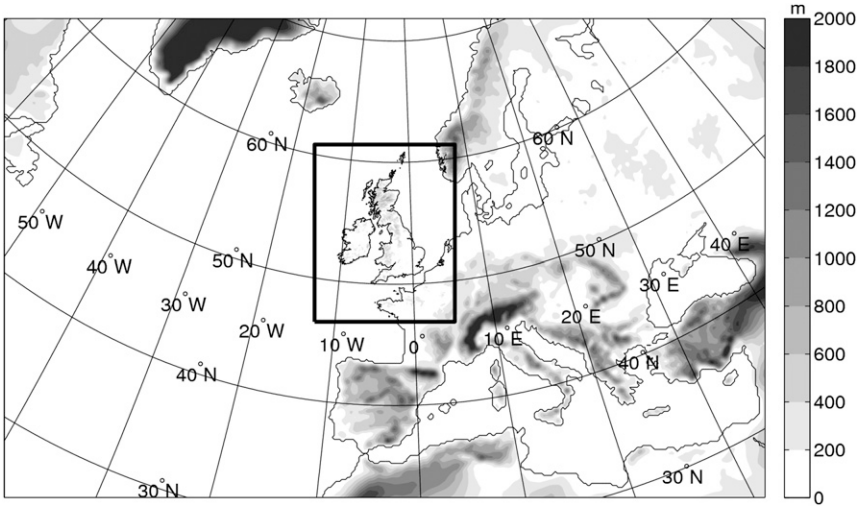

FIG. 3. Model domains for MOGREPS-R simulations (whole figure) and the UKV and MOGREPS-UK domain (inner rectangle). The shading shows the model terrain height for the respective models and the contours show the model coastlines.

convection-permitting models. Measures such as the root-mean-square (RMS) difference between predicted and observed fields become problematic at high resolution because the model is heavily penalized for misplacing the precipitation (Baldwin et al. 2001; Mass et al. 2002; Roebber et al. 2004). In fact, a small offset in the location of a convective cell could result in penalties being applied both where rain was observed but none was predicted and where rain was predicted but not observed (the so-called ''double penalty'' problem).

These failings of traditional verification measures have led to the creation of other verification methods for high-resolution forecasts (Ebert 2008). These can be broadly classified into four categories: neighborhood, scale separation, object based, and field deformation (Gilleland et al. 2009). The neighborhood-based approach compares data from a neighborhood of points rather than single grid points. It can involve an average of the precipitation totals over the neighborhood or the fraction of points within the neighborhood that exceed some threshold (Roberts and Lean 2008). Scale-separation approaches determine the properties of the precipitation based on their horizontal scale. The forecast is evaluated by applying a filter to the data [e.g., wavelets (Briggs and Levine 1997) or Fourier transforms (Harris et al. 2001)], which is used to quantify the scale-dependence of errors and the scale at which skill is lost. The field-deformation technique determines some optimal deformation to be applied to the forecast field to make it as similar as possible to the observed field. The field-deformation vectors representing the optimal deformation serve to quantify the differences between the fields (Hoffman et al. 1995; Alexander et al. 1999; Keil and Craig 2009). The object-based approach typically uses some threshold to identify individual objects in the instantaneous or accumulated precipitation fields and compares the statistics of the predicted and observed objects. One example of this approach is the structure–amplitude–location (SAL) technique (Wernli et al. 2008), where the structure, amplitude, and location of the objects in the forecast field are compared to those in the observed field.

Given that banded convection is associated with a particular elongated shape, the most obviously applicable verification method is object based. Previously, we used the SAL technique to evaluate the forecast skill for one convection-permitting ensemble (Barrett et al. 2015). However, this technique does not provide direct information on the timing or persistence of the convection, which are important characteristics of quasi-stationary bands. No existing verification method is able to simultaneously evaluate the object's structure, intensity, location, timing, and duration, though existing methods have been adapted to incorporate the timing aspect (Clark et al. 2014). Barrett et al. (2015) evaluated the stationarity of the bands subjectively by inspecting animations of the precipitation field, but such a manual approach is impractical for multiple ensembles. Hence, we have developed an extension of SAL that incorporates both timing and duration components, which is described below.

a. SAL verification

SAL quantifies differences in the structure (size and intensity), amplitude (total precipitation amount), and location of precipitation objects between forecast

and observed precipitation fields. These three components are described in detail below.

1) STRUCTURE COMPONENT

The structure component quantifies differences in the size and intensity of precipitation objects. Objects are identified in the precipitation field using a threshold. Wernli et al. (2008) suggest a threshold of 1/15 of the maximum precipitation rate in the verification box, but we use a threshold of 0.25 mm h^{-1} for all cases for consistency. A scaled volume V is then calculated as

$$V = \frac{\sum_{n=1}^{N} R_n^2 / R_n^{\max}}{\sum_{n=1}^{N} R_n}, \tag{1}$$

where R_n is the area-integrated precipitation of the object n and $R_n^{\max}$ is the peak precipitation value of any pixel within object n. The S component is then calculated as

$$S = \frac{V_{\text{model}} - V_{\text{obs}}}{0.5(V_{\text{model}} + V_{\text{obs}})}, \tag{2}$$

which falls between −2 and 2. Positive scores indicate that the model-simulated precipitation objects are either too large or have too low peak intensity.

2) AMPLITUDE COMPONENT

The amplitude component, which evaluates the verification-region-averaged precipitation, is calculated using

$$A = \frac{P_{\text{model}} - P_{\text{obs}}}{0.5(P_{\text{model}} + P_{\text{obs}})}, \tag{3}$$

where P is the verification-region-integrated precipitation. Positive scores indicate that the simulated precipitation is greater than observed. The range of A is also −2 to 2.

3) LOCATION COMPONENT

The location component L quantifies the physical distance between the centers of mass of the observed and model precipitation fields. Location L is composed of two components ($L = L_1 + L_2$): L_1 quantifies the distance between the verification-region centers of mass and L_2 quantifies the spread of objects around the verification-region center of mass. Component L_1 is calculated as

$$L_1 = \frac{|\mathbf{x}_{\text{model}} - \mathbf{x}_{\text{obs}}|}{d}, \tag{4}$$

where $\mathbf{x}$ is the center of mass of the precipitation field and d is the greatest distance between any two points in the verification box. For L_2, the distribution of precipitation objects around the center of mass is calculated as a weighted average distance between the centers of mass of the individual objects and the verification box center of mass, given by

$$r = \frac{\sum_{n=1}^{N} R_n |\mathbf{x}_{\text{domain}} - \mathbf{x}_n|}{\sum_{n=1}^{N} R_n}, \tag{5}$$

where $\mathbf{x}_{\text{domain}}$ and $\mathbf{x}_n$ are the centers of mass over the verification box and over object n, respectively. Component L_2 is then calculated as

$$L_2 = 2\left(\frac{|r_{\text{model}} - r_{\text{obs}}|}{d}\right). \tag{6}$$

Each component has a possible range of 0–1, giving an L range of 0–2. However, a score of 2 can never be achieved in practice because L_1 and L_2 are not independent and cannot both be large simultaneously. In general, larger L scores indicate that the simulated and observed precipitation centers of mass are farther apart or that the spread of the precipitation field around the verification-region center of mass is increasingly erroneous.

b. Extended SAL verification

Although SAL is typically applied to cumulative or instantaneous precipitation fields, it can be extended to evaluate the time evolution of the simulated precipitation in a similar framework. We do so herein by creating a Hovmöller plot (Hovmöller 1949) of precipitation within the verification box surrounding the observed band. The precipitation rate within the box is averaged in the cross-band direction and evaluated as a function of along-band distance and time. The grid interval of the Hovmöller plot is 1 km in along-band distance and 5 min in time. We apply a similar method to that followed in the calculation of the SAL L component to the Hovmöller to provide insight on both the position and timing of the precipitation. A third component is added to assess the precipitation duration.

1) STRUCTURE COMPONENT

The structure component is mathematically identical to that in the standard SAL, but applied to the Hovmöller plot. As before, a threshold of 0.25 mm h^{-1} is used for object detection.

2) AMPLITUDE COMPONENT

The amplitude component gives identical scores as in SAL if the same space and time verification boxes are considered and so is redundant if the standard SAL diagnostics are also calculated.

3) POSITION AND TIMING COMPONENTS

The position P and timing T components together describe the placement of the precipitation on the Hovmöller plot. The position component differs from SAL L component in three ways: 1) it is only applied to the along-wind-distance dimension of the Hovmöller plot, 2) it can have either a positive or negative value depending on whether the model center of mass is upwind or downwind of the observed center of mass, and 3) only the center of mass in the model and observed precipitation fields are compared. The timing component is identical to the position component except that it is applied to the time dimension of the Hovmöller plot. These components are given by

$$P = 2\left(\frac{x_{\text{model}} - x_{\text{obs}}}{d}\right), \quad \text{and} \tag{7}$$

$$T = 2\left(\frac{y_{\text{model}} - y_{\text{obs}}}{t}\right), \tag{8}$$

where (x, y) is the position of the center of mass along the (distance, time) axes, d is the length of the distance axis, and t is the timespan of the time axis. Values for P and T cannot be calculated if the model fails to produce any precipitation above the threshold rate.

The P and T components are equivalent to the L_1 component of SAL but with the modulus function removed, so as to provide information about whether the precipitation center of mass was upstream or downstream of the observed position and early or late. Positive P scores indicate that the forecast precipitation is farther downstream than observed and negative scores indicate that it is farther upstream. Similarly, the timing component evaluates whether the precipitation formed earlier ($T < 0$) or later ($T > 0$) than observed. No equivalent to L_2 has been included, so the normalized differences are scaled by a factor of 2 in the calculation of the P and T scores to achieve a range (from -2 to 2) that is consistent with other components.

4) DURATION COMPONENT

The duration component quantifies the persistence of precipitation in the Hovmöller plot. Although it has no equivalent SAL component, it is calculated similarly to A. Precipitation persistence is quantified at each point along the distance dimension of the Hovmöller plot as the number of pixels that exceed the threshold precipitation rate. The maximum number of pixels over all the locations M is taken to represent the persistence over the verification box. The values are calculated separately for model and observations and then compared to give the duration component D:

$$D = \frac{M_{\text{model}} - M_{\text{obs}}}{0.5(M_{\text{model}} + M_{\text{obs}})}. \tag{9}$$

c. *Interpretation of SAPTD scores*

Scores of each of the structure, amplitude, position, timing, and duration (SAPTD) components fall into the -2 to 2 range, which is identical to that of the structure and amplitude components of SAL but not the location component (0–2). In all cases, zero constitutes a perfect forecast of that component. For ease of reference, the physical significance of each of the components of SAL and SAPTD is summarized in Table 1.

4. Illustrative example: Case 1, 26 August 2012

To demonstrate the utility of the metrics defined above, we present them for case 1, which had the longest-lasting precipitation band and the largest localized precipitation accumulation of the four cases (Fig. 4a). The deterministic forecast initialized at 0900 UTC the previous day ($t - 18$ forecast, Fig. 4b) predicts a precipitation band that matches radar observations reasonably well in its location and alignment. However, the simulated band produced too little precipitation in its central region (from $-2°$ to $-3°$E) and too much precipitation both upwind and downwind of that region.

Some of the $t - 18$ MOGREPS-UK ensemble members also show banded precipitation accumulations that are reasonably consistent with the observations in some aspects (Fig. 5). These include the unperturbed "control" member (member 0) and members 5 and 9. However, either the location or orientation is not predicted correctly in these members. Other ensemble members display a variety of behaviors but generally fail to accurately reproduce the observed pattern.

The radar Hovmöller plot (Fig. 4d) shows that light precipitation fell almost continually between 75 and 125 km along the verification box, sometimes extending out to 160 km. Although the band was quasi-stationary and coherent, the Hovmöller plot reveals that the band contains embedded cells that travel downstream along its axis [consistent with the bands over southern France observed by Miniscloux et al. (2001)], as reflected by the embedded diagonal stripes. However, a clear anchoring point exists at ~75 km where precipitation is repeatedly initiated throughout the event.

TABLE 1. Interpretation of the different components of SAL and SAPTD.

Parameter	Negative scores	Positive scores
SAL: Structure	Precipitation covers too narrow an area, or peak accumulation value too high	Precipitation covers too broad an area, or peak accumulation value too low
SAL: Amplitude	Too little precipitation over verification box	Too much precipitation over verification
SAL: Location	Impossible	Larger implies greater separation of model and radar centers of mass or increasingly wrong spread about the radar centers of mass
SAPTD: Structure	Precipitation covers too little space in distance–time plot, or largest precipitation rate too high (small, intense cells)	Precipitation covers too much space in distance–time plot, or largest precipitation rate too low (broad, weaker precipitation)
SAPTD: Amplitude	Same value as SAL: Amplitude	Same value as SAL: Amplitude
SAPTD: Position	Precipitation center of mass too far upstream	Precipitation center of mass too far downstream
SAPTD: Timing	Precipitation center of mass too early	Precipitation center of mass too late
SAPTD: Duration	Precipitation duration too short	Precipitation duration too long

In comparison, the Hovmöllers of both the deterministic simulation (Fig. 4e) and the individual ensemble members (Fig. 6) reveal that the banded accumulations are not the result of a quasi-stationary band (as was observed) but rather isolated cells traversing the axis of the band. In the deterministic simulation, the large accumulation over the upwind sea results from persistent precipitation between 2200 and 0200 UTC, which reflects an upstream shift in the anchoring point. Although too intense, the precipitation over the downwind side of the verification box exhibits a similar timing as the observations (2200–0000 UTC).

The banded accumulation in the unperturbed member (member 0) also largely results from a mobile cell rather than a quasi-stationary band. However, quasi-stationary precipitation was apparent toward the end of the period (0700–0900 UTC) at a distance of ~75 km. The ensemble members predict precipitation much earlier in the period, on average, than was observed. The members generally fail to produce across-band-averaged rain rates above 0.25 mm h^{-1} during the period when the observed band was the most persistent (0000–0700 UTC).

Although one might expect the ensemble mean to provide useful information in quasi-stationary precipitation events (due to the fixed location of the precipitation), here it offers limited predictive value. The ensemble-mean precipitation accumulation (Fig. 4c) substantially differs from that observed (Fig. 4a) and that in the deterministic forecast (Fig. 4b). The ensemble mean shows a broad area of light precipitation due to the averaging of small features with disparate locations and timings. For the same reasons, the ensemble mean Hovmöller plot (Fig. 4f) differs from the radar and deterministic Hovmöllers (Figs. 4d,e) in its relative rarity of mean precipitation rates over 0.25 mm h^{-1}.

Interpretation of SAL and SAPTD scores

The standard SAL diagnostic scores (Fig. 7a) reveal that the individual $t-18$ ensemble members substantially underpredict the total precipitation accumulation over the verification box. The median amplitude score is -0.93 and no individual ensemble member predicts as much precipitation as was observed ($A<0$ for all members). The S score is generally positive, indicating that the precipitation is spread over too large an area or lacks the peak local accumulations. In the three ensemble members for which $S<0$, the total precipitation amount is grossly underpredicted. The L scores range from 0 to 0.5, suggesting that the location of the precipitation is reasonably well predicted across the ensemble. The smallest L scores correspond to ensemble members with larger A and larger S scores, reflecting broader precipitation objects that are approximately centered in the verification box.

The SAPTD diagnostics complement the SAL diagnostics by evaluating the time evolution of the simulated precipitation. Three of the four unique SAPTD components (S, T, and D) are shown in Fig. 7b, with the remaining component presented in the verification overview of Fig. 8a. The SAPTD D is strongly negative for the ensemble but slightly less negative for the deterministic simulation, which reflects the lack of persistent simulated precipitation in the Hovmöller plots (Fig. 4). Similarly, the SAPTD S scores are generally negative, consistent with D in that the objects in the Hovmöller plot are generally small and transient with larger precipitation rates than those observed. This

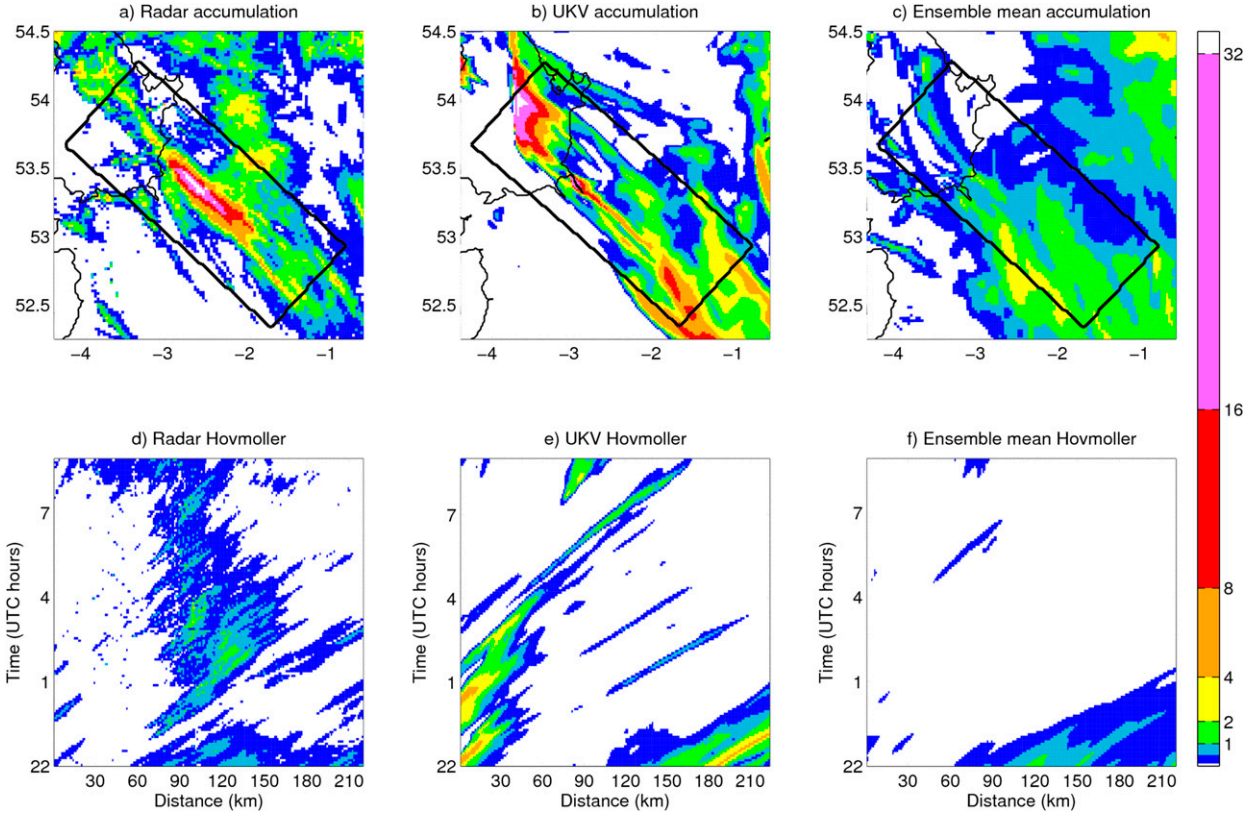

FIG. 4. (a)–(c) Accumulation (mm) and (d)–(f) Hovmöller plots (mm h^{-1}) for the 26 Aug 2012 case from radar measurements, UKV deterministic forecast initialized at 0900 UTC the previous day, and the MOGREPS-UK ensemble mean forecast initialized at 0900 UTC the previous day. Time period for all plots is 2200 UTC 25 Aug–0900 UTC 26 Aug 2012.

bias suggests that the precipitation accumulation results from a few intense, translating cells rather than stationary bands. The ensemble P scores are clustered around zero but slightly positive on average (Fig. 8a), revealing a trend for the precipitation to be shifted downstream.

The diagnostic scores for the ensemble and deterministic simulations initialized 6 h later ($t-12$) are shown in Figs. 7c and 7d. These simulations predict substantially more precipitation in general, with the median A increasing to -0.35 and four ensemble members predicting more precipitation than was observed over the verification box. The SAL S scores are again positive with a median of 0.94, indicating that the precipitation was spread over too large an area or lacked the peak local accumulations that were observed. The deterministic simulation is no longer an outlier from the set of ensemble members, but rather lies toward the middle of the S and A distributions. The L scores also decrease slightly, indicating better location accuracy in the later ensemble.

The SAPTD scores highlight that the precipitation is present for longer in the $t-12$ ensemble than in the $t-18$ ensemble (less negative D scores) and that the objects are broader and less intense (on the Hovmöller plot) than before (increased S scores), with 11 of the 12 members exhibiting S close to or above zero. The $t-12$ deterministic simulation scores quite similarly to the $t-18$ deterministic simulation for structure and duration, but the precipitation has moved slightly upstream (more negative P score) and occurs later (more positive T score).

Overall, the $t-12$ ensemble predicts more precipitation (larger SAL A) over a broader area (larger SAL S) than the $t-18$ ensemble. This increased precipitation results from more persistent events (larger SAPTD D) that are larger in scale in distance–time space (larger SAPTD S) than the isolated cells moving through the verification box in the $t-18$ ensemble. An example of these larger cells in the $t-18$ ensemble is shown in Figs. 9c and 9f. Comparing these cells and those of the $t-12$ ensemble (Figs. 9b,e) to the observed precipitation band (Figs. 9a,d), the precipitation morphology differs between the simulations and the observations.

5. Ensemble verification for all cases

The three remaining cases are now analyzed and results from all four cases summarized. Precipitation

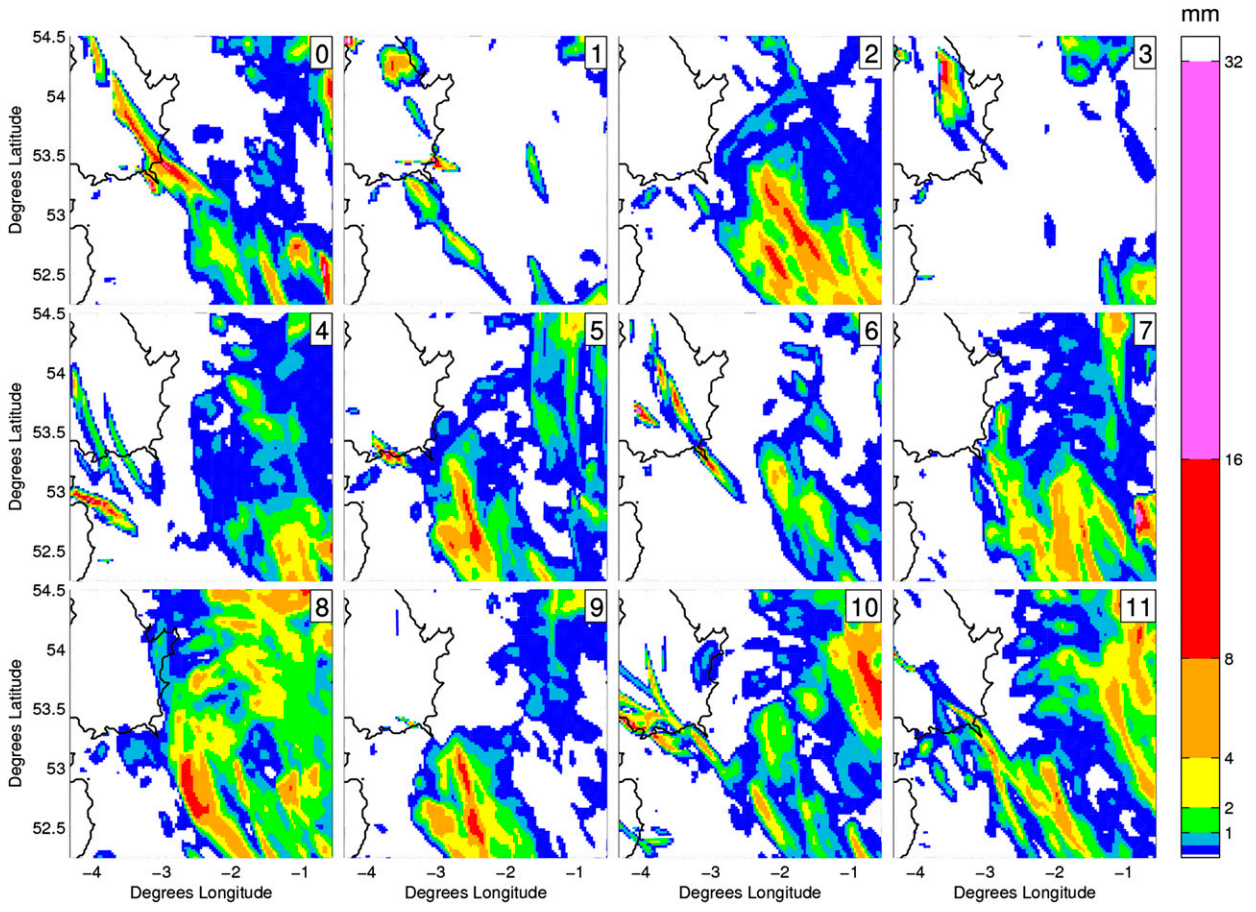

FIG. 5. Model precipitation accumulations for the 26 Aug 2012 case for the 12 ensemble members initialized at 0900 UTC the previous day. Precipitation from the 11-h period 2200 UTC 25 Aug–0900 UTC 26 Aug 2012 is shown, consistent with Figs. 4 and 6. The ensemble member number is marked in the upper-right corner of each panel.

accumulation and Hovmöller plots are shown for each case for both the radar and the best ensemble member from the $t-18$ ensemble (subjectively chosen based on the precipitation structure in both accumulation and Hovmöller plots). Diagnostic scores from SAL and SAPTD are summarized for both $t-18$ and $t-12$ simulations of each case in Fig. 8.

a. 28 August 2012

The most accurate member in the $t-18$ ensemble (Fig. 10b) exhibits much heavier and more widespread precipitation than that observed (Fig. 10a). These errors may arise from insufficient grid resolution (causing the individual cells to be too large) and insufficient convective inhibition in the impinging flow. In contrast to the observations where the convection initiated over land (Fig. 10c), the simulated precipitation initiates over the sea (perhaps because of insufficient convective inhibition there) before traversing the Welsh mountains (Fig. 10d). As in case 1, the simulations generally favor translating cells over the quasi-stationary bands that were observed. For the two ensembles as a whole, the SAL S and A scores are both spread around zero and include some extreme values, with extrema smaller in the $t-12$ ensemble than in the $t-18$ ensemble (Figs. 8c,d). The SAPTD D score is negative for every ensemble member, again reflecting that the model produces isolated and mobile convective cells rather than quasi-stationary bands. The precipitation centroids appear to be well represented in both ensembles, with SAPTD P values clustered around zero.

b. 9 September 2012

The radar-derived precipitation accumulation shows a broad area of precipitation with a flow-parallel streak of maximum accumulation through the middle of the verification box resulting from a persistent embedded convective band (Fig. 11a). The precipitation is located mostly over the higher terrain, concentrated mainly between 60 and 120 km distance on the Hovmöller plot

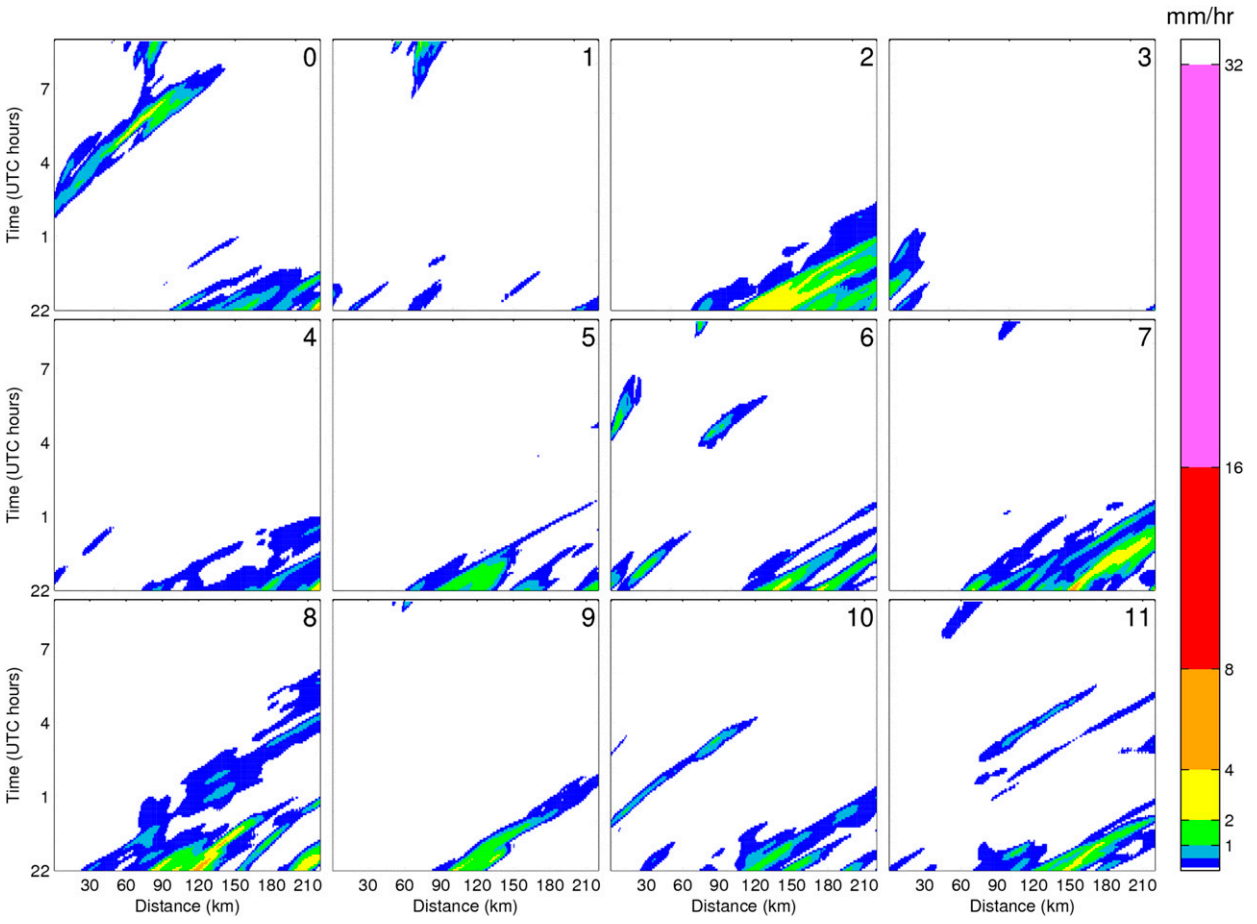

FIG. 6. Hovmöller plots of across-band averaged precipitation rate for the 26 Aug 2012 case for the 12 ensemble members initialized at 0900 UTC the previous day. The ensemble member number is marked in the upper-right corner of each panel.

(Fig. 11c). The Hovmöller plot from the selected $t-18$ ensemble simulation resembles the observations in the persistence of precipitation but overestimates its coverage and intensity (Figs. 11b,d). The diagnostic scores are tightly clustered for most of the SAL and SAPTD components in the $t-18$ ensemble and even tighter in the $t-12$ ensemble (Figs. 8e,f), suggesting a relatively small ensemble spread.

Although the ensemble members agree well with each other, they disagree with the observations in that their precipitation was too heavy (SAL $A>0$) and covered too large an area (SAL $S>0$). All of the ensemble members fail to reproduce the banded precipitation accumulations in the center of the verification box. This error is, by design, not captured by the SAPTD method because of its cross-flow averaging procedure. Although a persistent vertical stripe exists on the model Hovmöller plots, such a feature can correspond to either a broad area of stratiform precipitation or a quasi-stationary precipitation band. Because the model rainfall accumulations lack a banded structure, exhibit little intermittency, and correlate strongly with underlying terrain height (see Fig. 2), they are most likely owing to stratiform (rather than convective) orographic clouds. The absence of moist convection in the simulations suggests a systematic stable bias in the impinging flow.

Interestingly, the $t-12$ deterministic simulation is an outlier from the rest of the ensemble members in all of its verification scores (Fig. 8f). Although such consistently extreme behavior occurs only for this particular case and lead time, the deterministic simulation does have the most extreme score in 8 of the 49 other comparisons in Fig. 8.

c. 29 December 2012

This event gave rise to several bands of precipitation accumulation due to the combination of quasi-stationary bands over the Great Glen and embedded cells that translated through the verification box (Figs. 12a,c). Although the $t-18$ ensemble members exhibit general similarities with the observations, the band location

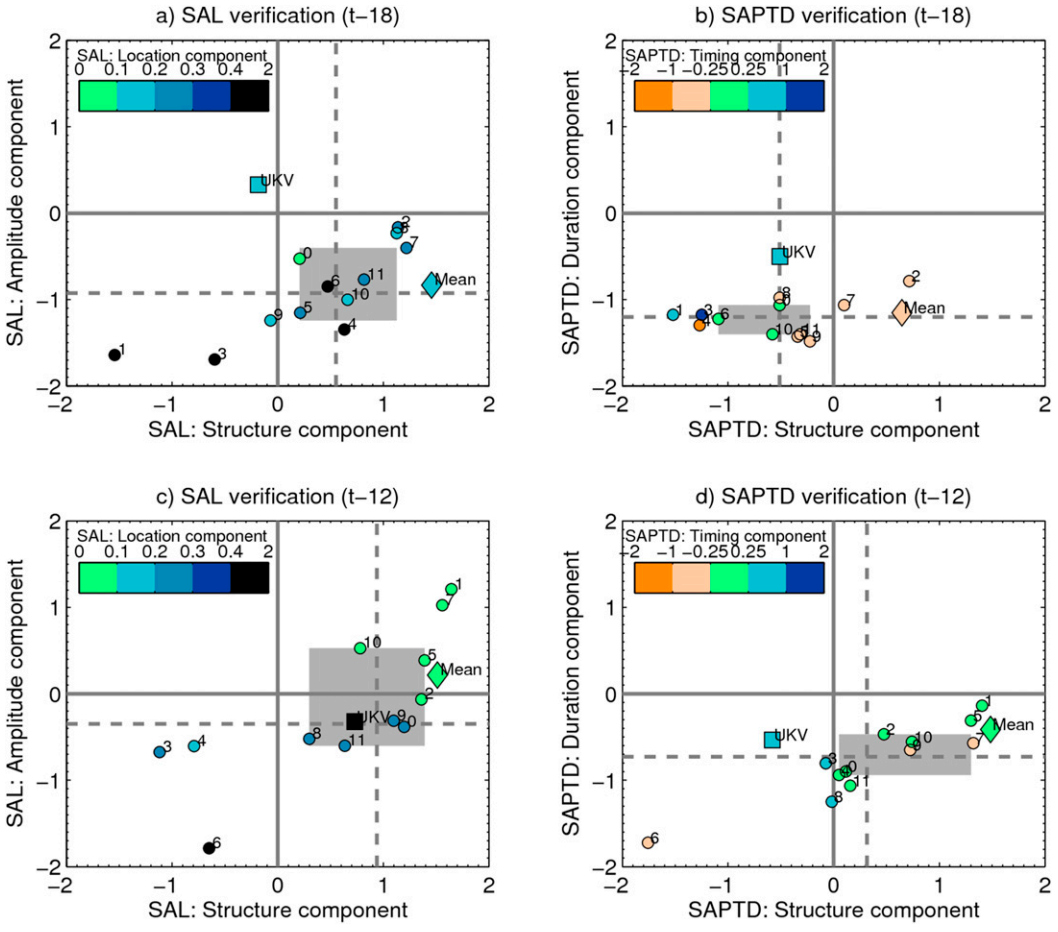

FIG. 7. (a),(c) SAL and (b),(d) SAPTD verification for the 26 Aug 2012 case using the ensemble initialized at (top) 0900 and (bottom) 1500 UTC the previous day. Each panel shows the individual ensemble members (circles), the deterministic simulation (squares), and the verification of the ensemble mean (diamonds). The ensemble-member median scores are marked by a dashed line, and the gray box denotes the interquartile range of these scores.

and stationarity varies greatly among them. The best ensemble member develops cells that initiate farther upstream, and translate downwind over a greater distance, than those in the observations (Figs. 12b,d). The ~30-km upstream shift in the band initiation, along with the coarse model representation of the convective cells, gives rise to an overly widespread precipitation accumulation that lacks the sharp cross-flow variability seen in the observations. As a result, the SAL scores for this member exhibit positive S and A (Figs. 8g,h). Other members of the $t-18$ ensemble exhibited similar structural and amplitude biases, as did most members of the $t-12$ ensemble (Fig. 8h).

Although the SAPTD S score is also large for all ensemble members, the D score is nearly zero for many of them. This combination suggests that the precipitation objects are more elongated in the model simulations than in observations, as found in the corresponding Hovmöller plots (Figs. 12c,d). These longer precipitation streaks correspond to discrete cells that initiated over the upstream sea and translated across the verification box.

d. Summary

The above four cases provide rich variability in the characteristic ensemble performance, along with some recurring themes. Although most members of the case-1 ensemble verify poorly against observations, the ensemble exhibits a large spread in SAL/SAPTD metrics. Apart from a notable underprediction in the duration of the simulated precipitation objects (as evidenced by a negative SAPTD D component), the spread of the ensemble straddles the SAPTD zero line in most metrics (Fig. 8), suggesting that the range of realizations broadly encompasses the observations. Similarly, the Case-2 simulations exhibit both large

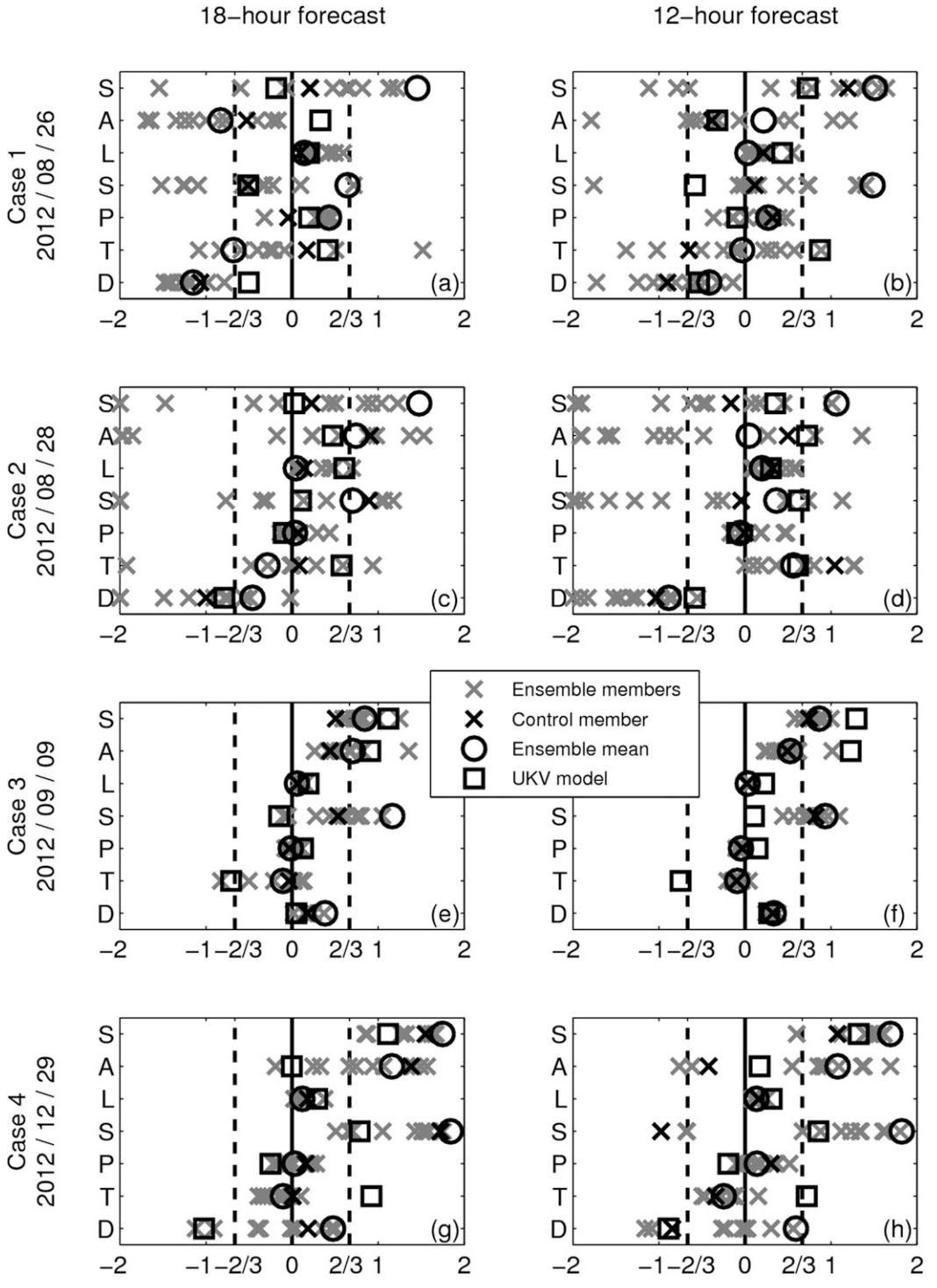

FIG. 8. Summary figure of SAL and SAPTD scores for each of the cases and both forecast lead times showing (left) the longest ($t - 18$) lead time and (right) the shortest ($t - 12$) lead time. Each row shows a different case. The first three rows of each plot show the components of SAL (structure, amplitude, and location), and the last four rows show components from SAPTD (structure, position, timing, and duration). Each row shows the individual ensemble members (gray crisscrosses), the unperturbed member (black crisscrosses), the mean precipitation from all ensemble members (not the mean verification score of the ensemble; black circles), and the higher-resolution forecast (black squares).

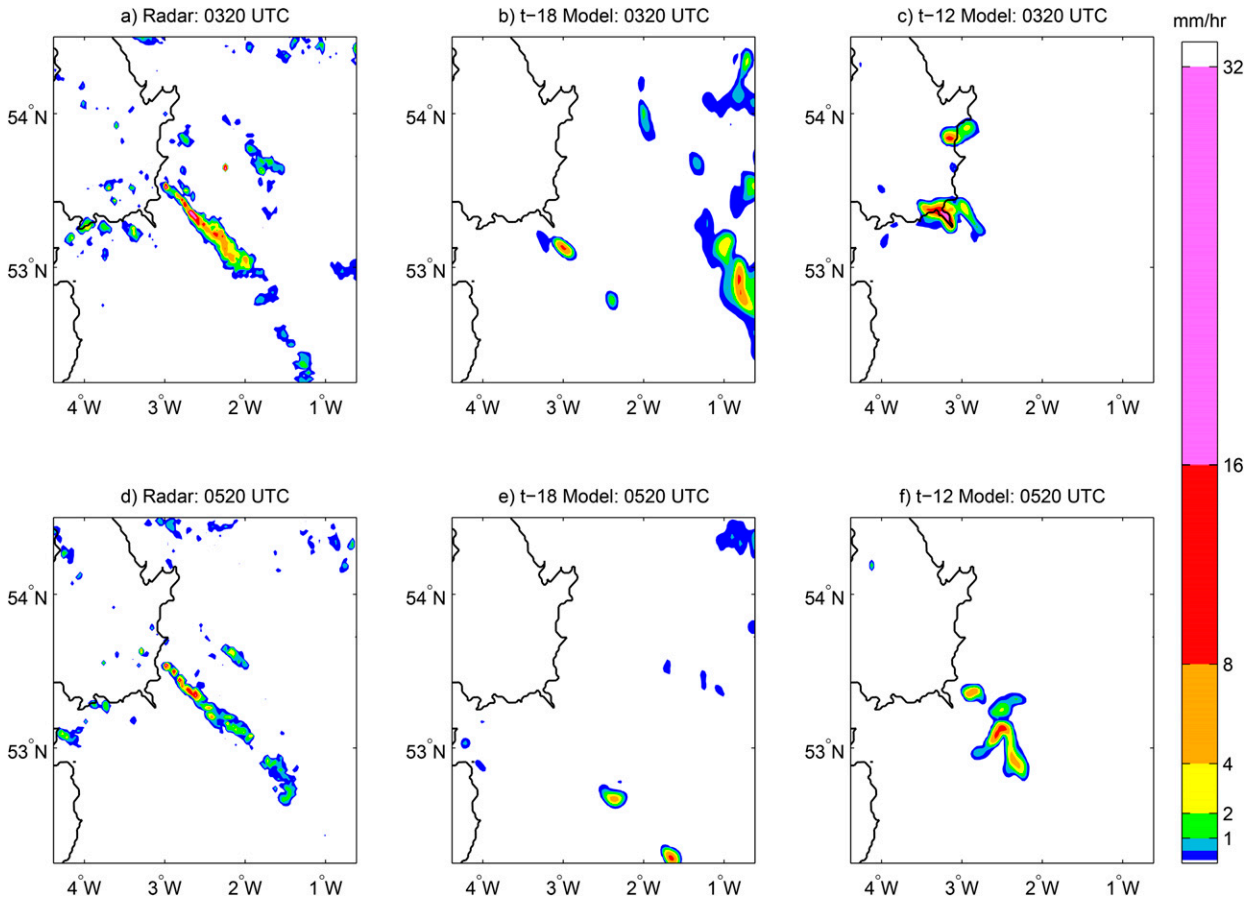

FIG. 9. Instantaneous rain rate fields at two times: (a)–(c) 0320 and (d)–(f) 0520 UTC on 26 Aug 2012, from the (a),(d) radar; (b),(e) a $t-18$ ensemble member; and (c),(f) a $t-12$ ensemble member.

error and large spread, with the spread again sufficient to cross the SAPTD zero line in all diagnostics except the D component. Case 3 exhibits a much smaller ensemble spread in the SAL/SAPTD metrics, but the ensemble is biased in the precipitation structure (too broad; $S>0$) and amplitude (too strong; $A>0$) and fails to reproduce the main convective band. Although the case-4 simulations reproduce the banded nature of the observed precipitation accumulations, the precipitation is again too broad and too strong.

With a small sample size of four cases, one cannot draw general conclusions about the MetUM ensemble skill. Nevertheless, we emphasize two fundamental features that the ensembles consistently struggle to represent: the persistence and the structure of the precipitation. The former is underestimated (SAPTD $D<0$) in three of the four cases (cases 1, 2, and 4). Although the model produces realistic bandlike precipitation accumulations, these accumulations result from discrete convective cells rather than the coherent, quasi-stationary bands that were observed (e.g., Figs. 4d, 9). Such errors may be owing to biases in the model representation of convective cells, in particular an inability to produce coherent convective bands when the environmental conditions favor them. A similar finding was obtained in 1.5-km MetUM simulations of a sea-breeze-forced convective band (Warren et al. 2014), which was corrected by reducing the grid spacing to 500 m to better resolve boundary layer circulations anchoring the convection. The one case that does not suffer from insufficient persistence (case 3) produces largely stable and stratiform precipitation, which diminishes the impact of errors owing to the representation of convection.

The structural bias common to all cases (SAPTD $S>0$) may be owing to a combination of limited model grid resolution and errors in the upstream flow conditions. The 1.5- and 2.2-km model grid spacings can only reasonably resolve processes with characteristic scales of about 10 km or larger. The convective bands of interest have typical widths of ~5 km or less (see Fig. 1) and are likely forced by terrain irregularities of

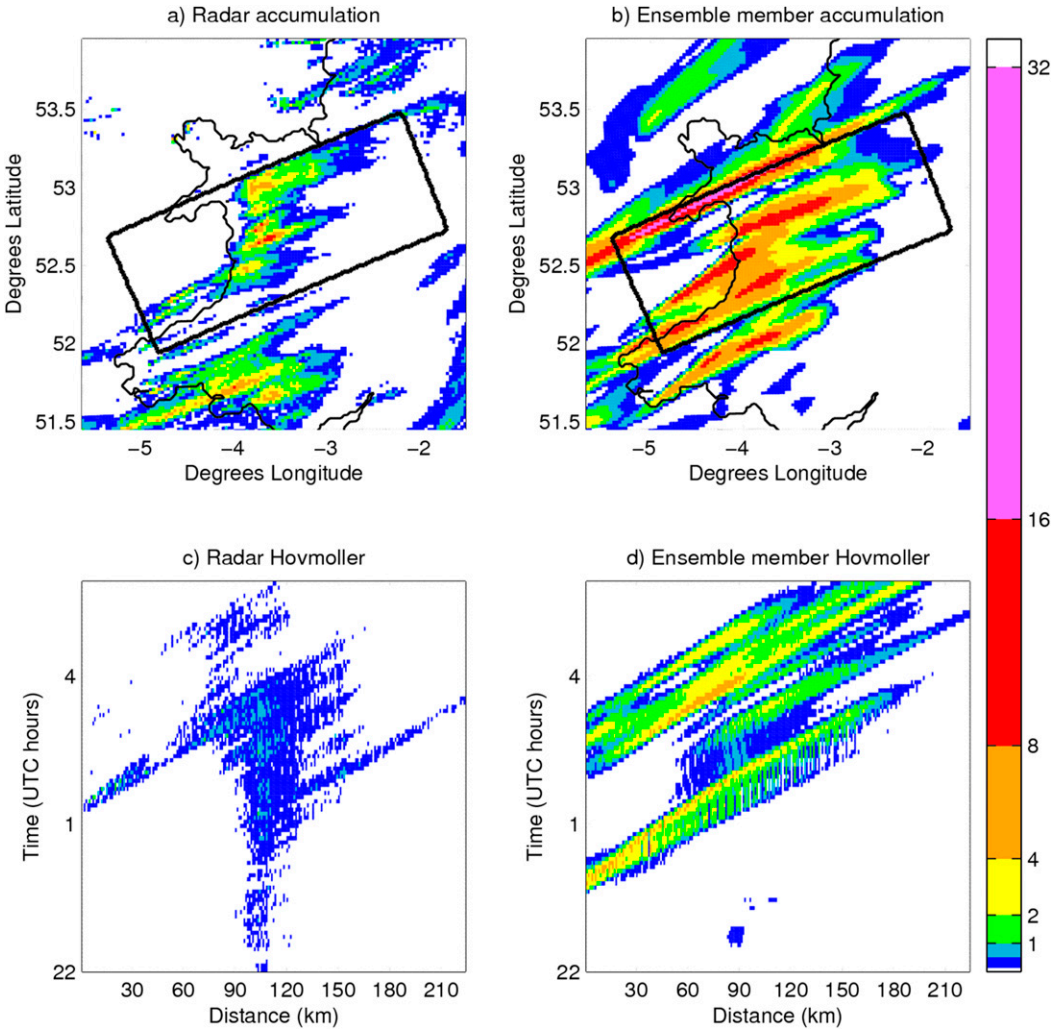

FIG. 10. The 8-h (a),(b) precipitation accumulation (mm) and (c),(d) Hovmöller plot (mm h^{-1}) from (left) radar and (right) best ensemble member for the 28 Aug 2012 case using the ensemble initialized at 0900 UTC the previous day.

the same scale, which renders them very poorly resolved. Although the quantitative impacts of this issue are minimized by SAPTD's cross-band averaging, it still compromises the representation of convective cells, the dominant precipitating features in cases 1, 2, and 4. In addition, errors in the upstream stability, moisture, and/or winds likely caused precipitation to initiate too far upstream and traverse a longer distance through the verification box, yielding larger and broader precipitation accumulations than those observed.

6. Enhanced-ensemble analysis

Because of the large computational cost associated with convection-permitting ensembles, it is important to maximize their value and to identify (and correct) the origins of their errors. The present section addresses these issues by considering, in turn, the third and fourth questions posed in section 1.

a. Is time lagging subsequent ensembles beneficial?

Based on the four cases under investigation and the seven verification diagnostics from SAL and SAPTD, we have evaluated the statistical properties of the ensembles. Of the 28 sets of verification scores (seven diagnostics over four cases), the mean verification score of the $t-12$ ensemble is reduced in magnitude compared to the $t-18$ ensemble in exactly half of the sets (14 of 28). The average scores are closer to zero in the $t-12$ ensemble by 0.5%, which is not statistically significant at the 95% confidence level.

Similarly, the ensemble spread was reduced in the $t-12$ ensemble in just over half (15 of 28) of the comparisons. The mean spread was 8% smaller for the $t-12$

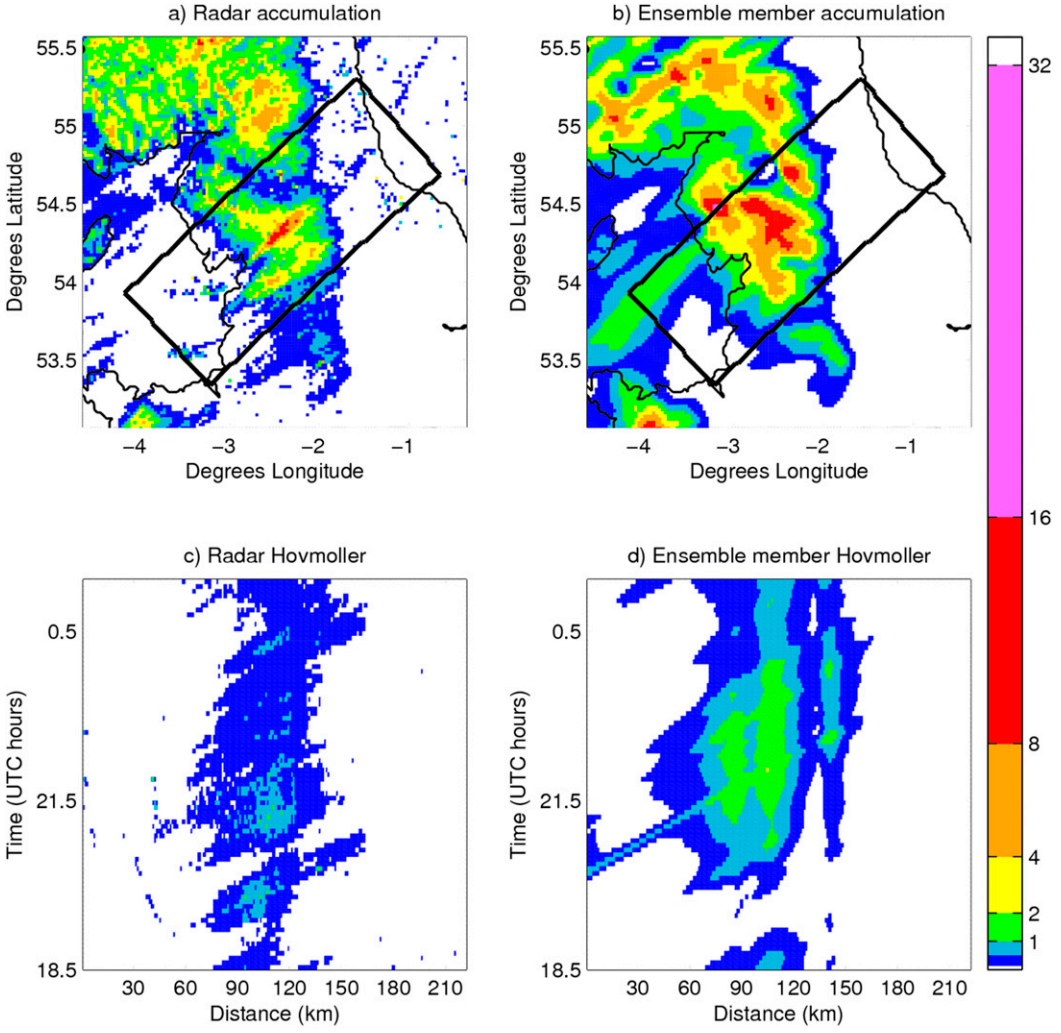

FIG. 11. The 6.5-h (a),(b) precipitation accumulation (mm) and (c),(d) Hovmöller plot (mm h^{-1}) from (left) radar and (right) best ensemble member for the 9 Sep 2012 case using the ensemble initialized at 0900 UTC the previous day.

ensemble, but again this is not statistically significant. Comparing the ensemble verification scores of subsequent ensemble runs, the mean scores are not statistically different for 24 of the 28 comparisons. The four cases where they do differ are for the SAL L and SAPTD D components for case 3 and the SAL S and A components for case 1.

A different perspective on the ensemble skill is provided by comparing relative operating characteristic (ROC) areas of the combined ensemble with the individual ensembles. The ROC verification is based on a 2×2 contingency table built from all forecast–observation pairs (here, the precipitation accumulation at every grid point within the verification box; e.g., Vié et al. 2011). With the false alarm rate (FAR) on the abscissa and probability of detection (POD) on the ordinate, the area under the ROC curve quantifies the skill of the ensemble in discriminating between events and nonevents. Table 2 presents comparisons of the area under the ROC curve for the four events, at different precipitation thresholds. In all cases and for almost all thresholds, the combined ensemble performs better than the worse ensemble (among $t-12$ and $t-18$) and often better than either ensemble alone.

Given that the SAPTD skill scores from successive ensembles are not significantly different herein, and that the ROC areas do not decrease by merging the two ensembles into one, it would be viable to combine the two ensembles to create an effective 24-member ensemble for these cases. Although we cannot generalize this result from our small sampling, these findings are

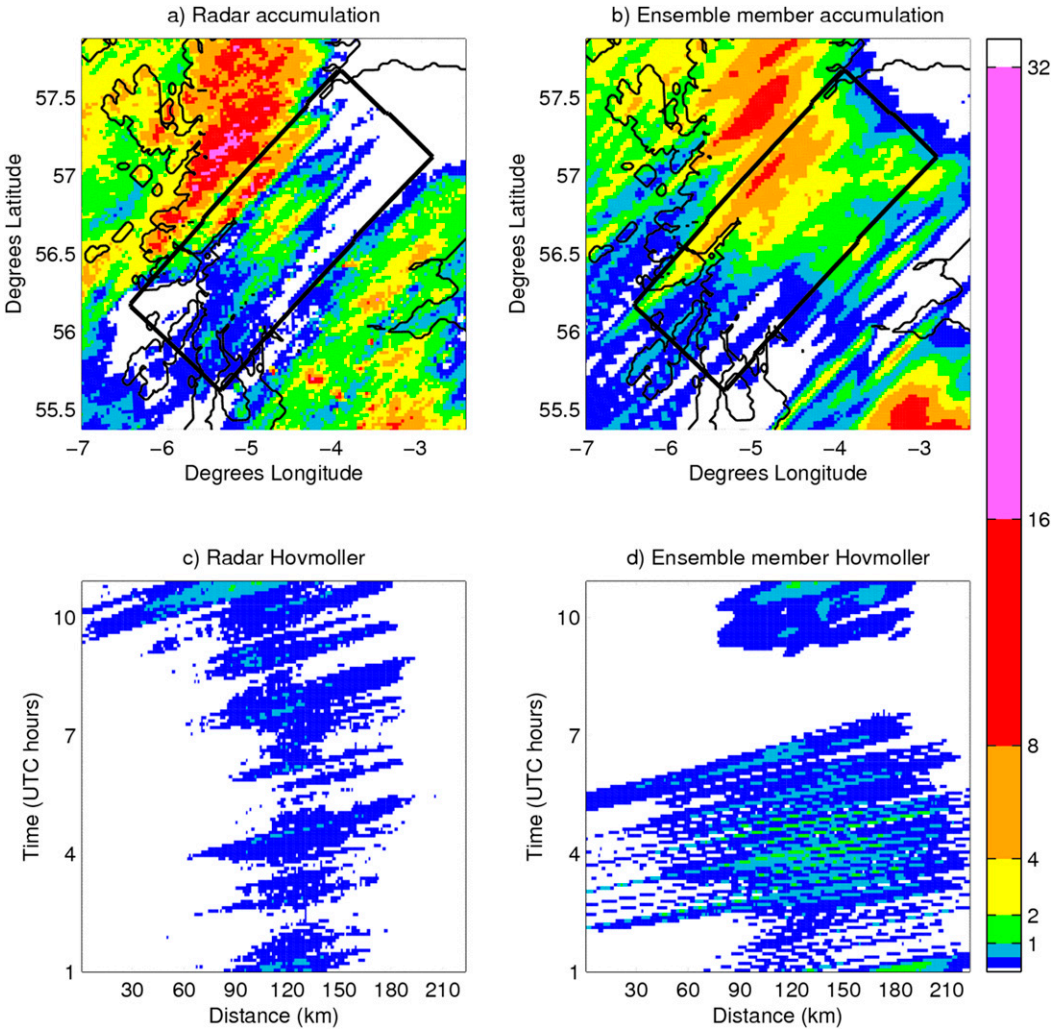

FIG. 12. The 10-h (a),(b) precipitation accumulation (mm) and (c),(d) Hovmöller plot (mm h^{-1}) from (left) radar and (right) best ensemble member for the 29 Dec 2012 case using the ensemble initialized at 0900 UTC the previous day.

consistent with Bouallègue et al. (2013), who time-lagged the German-focused Consortium for Small-Scale Modeling ensemble prediction system (COSMO-DE-EPS) ensemble simulations to improve the probability of precipitation forecasts by using a larger ensemble. They found that merging the most recent simulations with those from 3 and 6 h earlier, and equally weighting all ensemble members from all initialization times, provided a near-optimal solution. They also noted that time lagging could improve probabilistic precipitation forecasts without negatively impacting the temperature and wind forecasts through neighborhood averaging.

Combining subsequent ensemble sets into one larger ensemble may, for example, help to diminish spurious correlations in a convective-scale ensemble Kalman filter data assimilation system by increasing the statistical sampling. Similarly, one could merge the time-lagged ensembles for an ensemble sensitivity analysis to better isolate the initial and/or larger-scale factors that control the simulated band properties. If the skill of the ensembles are similar, then combining them does not necessarily produce a better forecast, but provides a larger sample from which to draw more robust statistics, as in the Bouallègue et al. (2013) study.

b. *What is the relationship between forecast accuracy at convective and larger scales?*

To examine the relationship between errors at different scales, we relate the errors' larger-scale flow to those in our nested ensembles by evaluating the mean errors in the simulated large-scale environment upstream of the observed bands. The larger-scale errors

TABLE 2. ROC areas calculated for all four cases using the $t-18$ and $t-12$ ensembles and the combined 24-member time-lagged ensemble. Values are calculated separately for each precipitation accumulation threshold. The boldface values highlight where the combined ensemble provides a better probabilistic forecast than either individual ensemble.

	Case 1			Case 2			Case 3			Case 4		
Threshold (mm)	$t-18$	$t-12$	Combined	$t-18$	$t-12$	Combined	$t-18$	$t-12$	Combined	$t-18$	$t-12$	Combined
0.25	0.6205	0.7328	0.6811	0.7319	0.6916	**0.7738**	0.8888	0.9344	**0.9344**	0.6569	0.5494	0.6038
0.5	0.6119	0.7280	0.6720	0.6838	0.6820	**0.7380**	0.9088	0.9367	**0.9370**	0.7673	0.6311	0.6958
1	0.6219	0.7364	0.6983	0.6643	0.6600	**0.6978**	0.9114	0.9336	0.9332	0.8420	0.6536	0.755
2	0.6031	0.6835	**0.6921**	0.6808	0.6742	**0.7188**	0.9025	0.9353	0.9316	0.8238	0.6300	0.7345
4	0.5552	0.6640	**0.6879**	0.6655	0.5214	0.6496	0.9334	0.9381	**0.9484**	0.6780	0.5915	0.6419
8	0.5004	0.6174	**0.6226**	0.4350	0.4241	0.3772	0.7603	0.9494	**0.9634**	0.6814	0.5022	0.6295
16	0.4982	0.5744	0.5726	—	—	—	—	—	—	0.5155	0.3894	0.4244
32	0.5000	0.4960	0.4960	—	—	—	—	—	—	—	—	—

are found by comparing the MOGREPS-R simulations with operational MetUM model analyses of pressure, temperature, and humidity at the surface (model level 1) and 5 km above the surface (near 500-hPa level; model level 27), along with eastward and northward wind components at 730 m above the surface (model level 10). These properties were evaluated over an area of 234 km × 198 km, located just upstream of the bands in all four cases. We then compute correlations between these errors and each of the components of SAL and SAPTD. The correlations were calculated separately for each case, giving 448 correlation coefficients in total.

Although intuition suggests that errors in the upstream flow are likely to physically correlate with errors in the orographic precipitation, some correlations may occur because of random chance. As we cannot easily determine which correlations are physically meaningful and which are statistically spurious, we assess whether the distribution of correlation coefficients differs from that produced if there was no physical relationship between the variables. To this end, we calculate correlation coefficients for 448 large-scale and convection-permitting pairs of 12-member ensembles containing randomly distributed values, and we repeat this process 1000 times.

The 1000 random samples of correlation coefficients clearly produce fewer strong correlations than those calculated from the model data (Fig. 13). For all correlation coefficients up to 0.85, the number of correlations from the model data exceeds that from the random surrogate data. For instance, there are 85 samples with correlation coefficients exceeding 0.5 compared to 42 ± 6 (one standard deviation) for the surrogate data. Therefore, the model data exceed the random threshold by seven standard deviations. For correlations exceeding 0.66, the difference is 10 standard deviations. These statistics reveal a significant correlation between the accuracy of the larger-scale forecast and the accuracy of simulated precipitation on the convection-permitting grids. Although the large-scale parameters that most strongly correlate with the simulated precipitation vary from case to case, the most influential large-scale parameters tend to be surface pressure and humidity and the near-surface wind speed (not shown).

7. Conclusions

This study has evaluated convection-permitting forecasts of terrain-locked and quasi-stationary convective bands forced by mesoscale topographic features over the United Kingdom. Four cases were selected based on analysis of data from the Met Office 1-km resolution radar network. In each case, a narrow precipitation band remained quasi-stationary for 1–7 h while producing moderate-to-high precipitation rates. Forecasts of these events from the Met Office Unified Model 2.2-km-grid-length ensemble and 1.5-km-grid-length deterministic model were verified against observations to quantify their skill in reproducing the observed precipitation.

The surface precipitation simulated by each of the models was verified using the Met Office radar-derived surface rain-rate product. To thoroughly evaluate the model performance, the structure, amount, timing, duration, and location of the precipitation were all compared to the observations. To facilitate such a comparison, we extended the SAL object-based verification method (Wernli et al. 2008) to apply it to distance–time (Hovmöller) plots (SAPTD). The main findings, which are separated into those pertaining to the model performance and those pertaining to the ensemble design, are summarized below.

a. *Model performance*

- The predictive skill of localized high-impact weather was highly variable among the four cases. The model was able to represent some aspects of location, intensity, structure, or duration of the precipitation in each case, but never all of them satisfactorily in a single case.

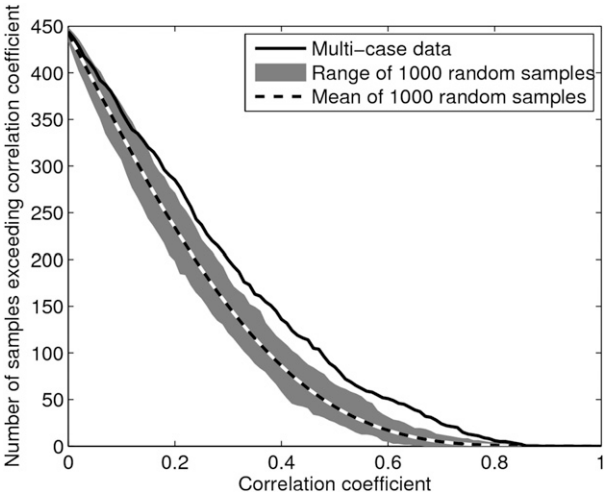

FIG. 13. Summary of the magnitude of the correlation coefficients from 448 pairs of 12-member ensemble data for the multicase data (black line) and from 1000 randomly generated sets of surrogate data. The mean (dashed line) and range (gray fill) from the surrogate data are shown.

- The stationarity of convection was inaccurately represented by the convection-permitting simulations, as reflected by underestimated precipitation-duration scores (the SAPTD *D* component) within the region of interest in three of the four cases. Although the model often produced bandlike precipitation accumulations that qualitatively resembled the observations, these resulted from a few mobile and intense convective cells rather than a quasi-stationary band.
- The accuracy of the model representation of the large-scale environment upstream of the band formation region was strongly correlated with the accuracy of the precipitation forecasts in the convection-permitting runs. The properties of the upstream environment that correlate the most strongly with the orographic precipitation varied from case to case but generally involved surface parameters.

b. *Ensemble design*

- In the four cases considered, the essential behavior of convection-permitting ensembles did not change significantly between subsequent ensemble cycles. Thus, artificially increasing the ensemble size by time-lagging subsequent ensemble cycles would have provided some benefits, including more confidence in the more predictable cases and larger ensemble spread in the less predictable cases.
- Consistent with previous studies of convection-permitting ensembles (e.g., Surcel et al. 2014), the ensemble-mean precipitation accumulation is more diffuse and lighter than that in the individual ensemble members. The structural characteristics of precipitating features in the individual members are lost when they are averaged over disparate locations and timings. Thus, metrics that retain the structure and intensity of the precipitation are required in order to provide warning of precipitation extremes when the location is less predictable.

Although the MetUM succeeds in producing banded precipitation accumulations, it generally struggles to develop or maintain elongated convective bands for the duration of the event. From such a small sampling we cannot determine whether the model systematically struggles to produce stationary bands or if it just fails to do so on these occasions. Moreover, we have not examined the false alarm rate for the model (i.e., how often it forecasts terrain-locked convective bands that are not observed). If these bands are subtly dependent on both the synoptic and local scale, as concluded by Barrett et al. (2015) and reinforced herein, the likelihood of any ensemble member correctly simulating the sequence of events producing the band would be small. Thus, a larger ensemble size may be needed to increase the likelihood of providing useful guidance on the possibility of such an event. Furthermore, the current operational model grid spacing is marginal for resolving these convective bands. Warren et al. (2014) found improvements in the simulation of a quasi-stationary convective band by reducing the grid spacing to 500 m. Thus, higher-resolution ensembles may be required to overcome the problems with band morphology.

Further work is required to determine whether convection-permitting ensembles such as MOGREPS-UK can accurately predict the detailed structure and propagation of convection more generally. Such research should verify the ensemble representation of both specific mesoscale phenomena (as was done here in the case of quasi-stationary bands) and varied phenomena over diverse events. For these purposes, methods like SAPTD will help to thoroughly quantify the model representation of both structural and temporal aspects of the convective precipitation.

Acknowledgments. This work has been funded through the Natural Environment Research Council (NERC) as part of the PRESTO (Precipitation Structures over Orography) project (NE/1024984/1). We thank the Met Office for making the MetUM available for research purposes and the National Centre for Atmospheric Science (NCAS) Computational Modelling Support (CMS) for providing technical support, in particular via Willie McGinty. We would also like to thank Humphrey Lean, Kirsty Hanley, Simon Vosper, and Paul Field for useful

discussions. We also thank Russ Schumacher and an anonymous reviewer for their constructive comments, which have helped to improve this manuscript. Model data from this project are archived by, and available from, the Centre for Environmental Data Analysis (www.ceda.ac.uk).

REFERENCES

Alexander, G. D., J. A. Weinman, V. M. Karyampudi, W. S. Olson, and A. Lee, 1999: The effect of assimilating rain rates derived from satellites and lightning on forecasts of the 1993 superstorm. *Mon. Wea. Rev.*, **127**, 1433–1457, doi:10.1175/1520-0493(1999)127<1433:TEOARR>2.0.CO;2.

Baldwin, M. E., S. Lakshmivarahan, and J. S. Kain, 2001: Verification of mesoscale features in NWP models. Preprints, *Ninth Conf. on Mesoscale Processes*, Fort Lauderdale, FL, Amer. Meteor. Soc., 255–258.

Barrett, A. I., S. L. Gray, D. J. Kirshbaum, N. M. Roberts, D. M. Schultz, and J. G. Fairman Jr., 2015: Synoptic versus orographic control on stationary convective banding. *Quart. J. Roy. Meteor. Soc.*, **141**, 1101–1113, doi:10.1002/qj.2409.

Bouallègue, Z. B., S. E. Theis, and C. Gebhardt, 2013: Enhancing COSMO-DE ensemble forecasts by inexpensive techniques. *Meteor. Z.*, **22**, 49–59, doi:10.1127/0941-2948/2013/0374.

Bowler, N. E., and K. R. Mylne, 2009: Ensemble transform Kalman filter perturbations for a regional ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **135**, 757–766, doi:10.1002/qj.404.

——, A. Arribas, K. R. Mylne, K. B. Robertson, and S. E. Beare, 2008: The MOGREPS short-range ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **134**, 703–722, doi:10.1002/qj.234.

——, ——, S. E. Beare, K. R. Mylne, and G. J. Shutts, 2009: The local ETKF and SKEB: Upgrades to the MOGREPS short-range ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **135**, 767–776, doi:10.1002/qj.394.

Briggs, W. M., and R. A. Levine, 1997: Wavelets and field forecast verification. *Mon. Wea. Rev.*, **125**, 1329–1341, doi:10.1175/1520-0493(1997)125<1329:WAFFV>2.0.CO;2.

Brown, A., S. Milton, M. Cullen, B. Golding, J. Mitchell, and A. Shelly, 2012: Unified modeling and prediction of weather and climate. *Bull. Amer. Meteor. Soc.*, **93**, 1865–1877, doi:10.1175/BAMS-D-12-00018.1.

Browning, K., A. Eccleston, and G. Monk, 1985: The use of satellite and radar imagery to identify persistent shower bands downwind of the North Channel. *Meteor. Mag.*, **114** (1360), 325–331.

Clark, A. J., W. A. Gallus, M. Xue, and F. Kong, 2009: A comparison of precipitation forecast skill between small convection-allowing and large convection-parameterizing ensembles. *Wea. Forecasting*, **24**, 1121–1140, doi:10.1175/2009WAF2222222.1.

——, W. A. Gallus Jr., M. Xue, and F. Kong, 2010: Convection-allowing and convection-parameterizing ensemble forecasts of a mesoscale convective vortex and associated severe weather environment. *Wea. Forecasting*, **25**, 1052–1081, doi:10.1175/2010WAF2222390.1.

——, and Coauthors, 2011: Probabilistic precipitation forecast skill as a function of ensemble size and spatial scale in a convection-allowing ensemble. *Mon. Wea. Rev.*, **139**, 1410–1418, doi:10.1175/2010MWR3624.1.

——, and Coauthors, 2012: An overview of the 2010 hazardous weather testbed experimental forecast program spring experiment. *Bull. Amer. Meteor. Soc.*, **93**, 55–74, doi:10.1175/BAMS-D-11-00040.1.

——, R. G. Bullock, T. L. Jensen, M. Xue, and F. Kong, 2014: Application of object-based time-domain diagnostics for tracking precipitation systems in convection-allowing models. *Wea. Forecasting*, **29**, 517–542, doi:10.1175/WAF-D-13-00098.1.

Clark, M. R., 2011: An observational study of the exceptional 'Ottery St Mary' thunderstorm of 30 October 2008. *Meteor. Appl.*, **18**, 137–154, doi:10.1002/met.187.

Cosma, S., E. Richard, and F. Miniscloux, 2002: The role of small-scale orographic features in the spatial distribution of precipitation. *Quart. J. Roy. Meteor. Soc.*, **128**, 75–92, doi:10.1256/003590002604987 98.

Davies, T., M. J. P. Cullen, A. J. Malcolm, M. H. Hawson, A. Staniforth, A. A. White, and N. Wood, 2005: A new dynamical core for the Met Office's global and regional modelling of the atmosphere. *Quart. J. Roy. Meteor. Soc.*, **131**, 1759–1782, doi:10.1256/qj.04.101.

Ebert, E. E., 2008: Fuzzy verification of high-resolution gridded forecasts: A review and proposed framework. *Meteor. Appl.*, **15**, 51–64, doi:10.1002/met.25.

Edwards, J. M., and A. Slingo, 1996: Studies with a flexible new radiation code. I: Choosing a configuration for a large-scale model. *Quart. J. Roy. Meteor. Soc.*, **122**, 689–719, doi:10.1002/qj.49712253107.

Elmore, K. L., S. J. Weiss, and P. C. Banacos, 2003: Operational ensemble cloud model forecasts: Some preliminary results. *Wea. Forecasting*, **18**, 953–964, doi:10.1175/1520-0434(2003)018<0953:OECMFS>2.0.CO;2.

Fairman, J. G., D. M. Schultz, D. J. Kirshbaum, S. L. Gray, and A. I. Barrett, 2015: A radar-based rainfall climatology of Great Britain and Ireland. *Weather*, **70**, 153–158, doi:10.1002/wea.2486.

Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of spatial forecast verification methods. *Wea. Forecasting*, **24**, 1416–1430, doi:10.1175/2009WAF2222269.1.

Golding, B., 2005: The Boscastle flood: Meteorological analysis of the conditions leading to the flooding on 16 August 2004. *Weather*, **60**, 230–235, doi:10.1256/wea.71.05.

Gregory, D., and P. Rowntree, 1990: A mass flux convection scheme with representation of cloud ensemble characteristics and stability dependent closure. *Mon. Wea. Rev.*, **118**, 1483–1506, doi:10.1175/1520-0493(1990)118<1483:AMFCSW>2.0.CO;2.

Hanley, K. E., D. J. Kirshbaum, S. E. Belcher, N. M. Roberts, and G. Leoncini, 2011: Ensemble predictability of an isolated mountain thunderstorm in a high-resolution model. *Quart. J. Roy. Meteor. Soc.*, **137**, 2124–2137, doi:10.1002/qj.877.

——, ——, N. M. Roberts, and G. Leoncini, 2013: Sensitivities of a squall line over central Europe in a convective-scale ensemble. *Mon. Wea. Rev.*, **141**, 112–133, doi:10.1175/MWR-D-12-00013.1.

Harris, D., E. Foufoula-Georgiou, K. K. Droegemeier, and J. J. Levit, 2001: Multiscale statistical properties of a high-resolution precipitation forecast. *J. Hydrometeor.*, **2**, 406–418, doi:10.1175/1525-7541(2001)002<0406:MSPOAH>2.0.CO;2.

Harrison, D. L., R. W. Scovell, and M. Kitchen, 2009: High-resolution precipitation estimates for hydrological uses. *Proc. Inst. Civ. Eng. Water Manage.*, **162**, 125–135, doi:10.1680/wama.2009.162.2.125.

——, K. Norman, C. Pierce, and N. Gaussiat, 2012: Radar products for hydrological applications in the UK. *Proc. Inst. Civ. Eng. Water Manage.*, **165**, 89–103, doi:10.1680/wama.2012.165.2.89.

Hoffman, R. N., Z. Liu, J.-F. Louis, and C. Grassoti, 1995: Distortion representation of forecast errors. *Mon. Wea. Rev.*, **123**, 2758–2770, doi:10.1175/1520-0493(1995)123<2758:DROFE>2.0.CO;2.

Hovmöller, E., 1949: The trough-and-ridge diagram. *Tellus*, **1**, 62–66, doi:10.1111/j.2153-3490.1949.tb01260.x.

Keil, C., and G. C. Craig, 2009: A displacement and amplitude score employing an optical flow technique. *Wea. Forecasting*, **24**, 1297–1308, doi:10.1175/2009WAF2222247.1.

Kirshbaum, D. J., and D. R. Durran, 2005: Observations and modeling of banded orographic convection. *J. Atmos. Sci.*, **62**, 1463–1479, doi:10.1175/JAS3417.1.

——, G. H. Bryan, R. Rotunno, and D. R. Durran, 2007a: The triggering of orographic rainbands by small-scale topography. *J. Atmos. Sci.*, **64**, 1530–1549, doi:10.1175/JAS3924.1.

——, R. Rotunno, and G. H. Bryan, 2007b: The spacing of orographic rainbands triggered by small-scale topography. *J. Atmos. Sci.*, **64**, 4222–4245, doi:10.1175/2007JAS2335.1.

Kong, F., K. K. Droegemeier, and N. L. Hickmon, 2006: Multiresolution ensemble forecasts of an observed tornadic thunderstorm system. Part I: Comparison of coarse-and fine-grid experiments. *Mon. Wea. Rev.*, **134**, 807–833, doi:10.1175/MWR3097.1.

——, ——, and ——, 2007: Multiresolution ensemble forecasts of an observed tornadic thunderstorm system. Part II: Storm-scale experiments. *Mon. Wea. Rev.*, **135**, 759–782, doi:10.1175/MWR3323.1.

Lean, H. W., P. A. Clark, M. Dixon, N. M. Roberts, A. Fitch, R. Forbes, and C. Halliwell, 2008: Characteristics of high-resolution versions of the Met Office Unified Model for forecasting convection over the United Kingdom. *Mon. Wea. Rev.*, **136**, 3408–3424, doi:10.1175/2008MWR2332.1.

Lock, A. P., A. R. Brown, M. R. Bush, G. M. Martin, and R. N. B. Smith, 2000: A new boundary layer mixing scheme. Part I: Scheme description and single-column model tests. *Mon. Wea. Rev.*, **128**, 3187–3199, doi:10.1175/1520-0493(2000)128<3187:ANBLMS>2.0.CO;2.

Mass, C. F., 1981: Topographically forced convergence in western Washington state. *Mon. Wea. Rev.*, **109**, 1335–1347, doi:10.1175/1520-0493(1981)109<1335:TFCIWW>2.0.CO;2.

——, D. Ovens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bull. Amer. Meteor. Soc.*, **83**, 407–430, doi:10.1175/1520-0477(2002)083<0407:DIHRPM>2.3.CO;2.

Miniscloux, F., J. D. Creutin, and S. Anquetin, 2001: Geostatistical analysis of orographic rainbands. *J. Appl. Meteor.*, **40**, 1835–1854, doi:10.1175/1520-0450(2001)040<1835:GAOOR>2.0.CO;2.

Mylne, K., 2013: Scientific framework for the ensemble prediction system for the UKV. MOSAC Paper 18.6, Met Office, 12 pp. [Available online at http://www.metoffice.gov.uk/media/pdf/q/0/MOSAC_18.6_Mylne.pdf.]

Novak, D. R., and B. A. Colle, 2012: Diagnosing snowband predictability using a multimodel ensemble system. *Wea. Forecasting*, **27**, 565–585, doi:10.1175/WAF-D-11-00047.1.

Richardson, D. S., 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, doi:10.1002/qj.49712757715.

Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, doi:10.1175/2007MWR2123.1.

Roebber, P. J., D. M. Schultz, B. A. Colle, and D. J. Stensrud, 2004: Toward improved prediction: High-resolution and ensemble modeling systems in operations. *Wea. Forecasting*, **19**, 936–949, doi:10.1175/1520-0434(2004)019<0936:TIPHAE>2.0.CO;2.

——, K. L. Swanson, and J. K. Ghorai, 2008: Synoptic control of mesoscale precipitating systems in the Pacific Northwest. *Mon. Wea. Rev.*, **136**, 3465–3476, doi:10.1175/2008MWR2264.1.

Smagorinsky, J., 1963: General circulation experiments with the primitive equations. *Mon. Wea. Rev.*, **91**, 99–164, doi:10.1175/1520-0493(1963)091<0099:GCEWTP>2.3.CO;2.

Smith, R. N. B., 1990: A scheme for predicting layer clouds and their water content in a general circulation model. *Quart. J. Roy. Meteor. Soc.*, **116**, 435–460, doi:10.1002/qj.49711649210.

Stensrud, D. J., and Coauthors, 2009: Convective-scale warn-on-forecast system: A vision for 2020. *Bull. Amer. Meteor. Soc.*, **90**, 1487–1499, doi:10.1175/2009BAMS2795.1.

Suarez, A., H. D. Reeves, D. Wheatley, and M. Coniglio, 2012: Comparison of ensemble Kalman filter-based forecasts to traditional ensemble and deterministic forecasts for a case study of banded snow. *Wea. Forecasting*, **27**, 85–105, doi:10.1175/WAF-D-11-00030.1.

Surcel, M., I. Zawadzki, and M. Yau, 2014: On the filtering properties of ensemble averaging for storm-scale precipitation forecasts. *Mon. Wea. Rev.*, **142**, 1093–1105, doi:10.1175/MWR-D-13-00134.1.

Tang, Y., H. W. Lean, and J. Bornemann, 2013: The benefits of the Met Office variable resolution NWP model for forecasting convection. *Meteor. Appl.*, **20**, 417–426, doi:10.1002/met.1300.

Vié, B., O. Nuissier, and V. Ducrocq, 2011: Cloud-resolving ensemble simulations of Mediterranean heavy precipitating events: Uncertainty on initial conditions and lateral boundary conditions. *Mon. Wea. Rev.*, **139**, 403–423, doi:10.1175/2010MWR3487.1.

Warren, R. A., D. J. Kirshbaum, R. S. Plant, and H. W. Lean, 2014: A 'Boscastle-type' quasi-stationary convective system over the UK southwest peninsula. *Quart. J. Roy. Meteor. Soc.*, **140**, 240–257, doi:10.1002/qj.2124.

Wernli, H., M. Paulat, M. Hagen, and C. Frei, 2008: SAL—A novel quantity measure for the verification of quantitative precipitation forecasts. *Mon. Wea. Rev.*, **136**, 4470–4487, doi:10.1175/2008MWR2415.1.

Wilson, R. W., and S. P. Ballard, 1999: A microphysically based precipitation scheme for the UK Meteorological Office Unified Model. *Quart. J. Roy. Meteor. Soc.*, **125**, 1607–1636, doi:10.1002/qj.49712555707.

Yoshizaki, M., T. Kato, Y. Tanaka, H. Takayama, Y. Shoji, and H. Seko, 2000: Analytical and numerical study of the 26 June 1998 orographic rainband observed in Western Kyushu, Japan. *J. Meteor. Soc. Japan*, **78** (6), 835–856.