Medium-Range Convection-Allowing Ensemble Forecasts with a Variable-Resolution Global Model

CRAIG S. SCHWARTZ

National Center for Atmospheric Research, Boulder, Colorado

(Manuscript received 20 December 2018, in final form 6 March 2019)

ABSTRACT

Two sets of global, 132-h (5.5-day), 10-member ensemble forecasts were produced with the Model for Prediction Across Scales (MPAS) for 35 cases in April and May 2017. One MPAS ensemble had a quasi-uniform 15-km mesh while the other employed a variable-resolution mesh with 3-km cell spacing over the conterminous United States (CONUS) that smoothly relaxed to 15 km over the rest of the globe. Precipitation forecasts from both MPAS ensembles were objectively verified over the central and eastern CONUS to assess the potential benefits of configuring MPAS with a 3-km mesh refinement region for medium-range forecasts. In addition, forecasts from NCEP's operational Global Ensemble Forecast System were evaluated and served as a baseline against which to compare the experimental MPAS ensembles. The 3-km MPAS ensemble most faithfully reproduced the observed diurnal cycle of precipitation throughout the 132-h forecasts and had superior precipitation skill and reliability over the first 48 h. However, after 48 h, the three ensembles had more similar spread, reliability, and skill, and differences between probabilistic precipitation forecasts derived from the 3- and 15-km MPAS ensembles were typically statistically insignificant. Nonetheless, despite fewer benefits of increased resolution for spatial placement after 48 h, 3-km ensemble members explicitly provided potentially valuable guidance regarding convective mode throughout the 132-h forecasts while the other ensembles did not. Collectively, these results suggest both strengths and limitations of medium-range high-resolution ensemble forecasts and reveal pathways for future investigations to improve understanding of high-resolution global ensembles with variable-resolution meshes.

1. Introduction

High-resolution, convection-allowing numerical weather prediction (NWP) models with horizontal grid spacings of approximately 4 km or less are ubiquitous in research and operations and produce better short-term precipitation and severe weather forecasts than coarser-resolution, convection-parameterizing NWP models (e.g., [Done et al. 2004](#); [Kain et al. 2006](#); [Roberts and Lean 2008](#); [Weisman et al. 2008](#); [Clark et al. 2009, 2010a](#); [Schwartz et al. 2009](#); [Duc et al. 2013](#); [Schwartz 2014](#); [Clark et al. 2016](#); [Iyer et al. 2016](#); [Schellander-Gorgas et al. 2017](#)). Because of computational constraints, convection-allowing models have traditionally been run solely over limited-area domains, meaning lateral boundary conditions (LBCs) provided by coarser external models eventually infiltrate domain interiors and limit skill; the speed at which LBCs contaminate a forecast depends on

domain size and meteorological regime (e.g., [Warner et al. 1997](#)).

Conversely, global models are free from LBC-related constraints and errors and are therefore attractive for longer-term forecasting. Moreover, increased computing resources have enabled global convection-allowing forecasts for retrospective case studies and idealized scenarios (e.g., [Miura et al. 2007](#); [Satoh et al. 2008, 2017](#); [Putman and Suarez 2011](#); [Miyamoto et al. 2013, 2016](#); [Skamarock et al. 2014](#); [Heinzeller et al. 2016](#); [Judt 2018](#)), although real-time global convection-allowing forecasts have not been attempted.

However, many global NWP models have the capability to refine horizontal grid spacing over a portion of the planet while maintaining coarser resolution elsewhere (e.g., [Fox-Rabinovitz et al. 2006](#); [Tomita 2008](#); [Skamarock et al. 2012](#); [Harris and Lin 2013](#); [Harris et al. 2016](#)), and these variable-resolution meshes can sufficiently reduce computational costs to facilitate real-time global forecasts with regions of convection-allowing resolution. Indeed, as contributions

Corresponding author: Craig Schwartz, schwartz@ucar.edu

DOI: 10.1175/MWR-D-18-0452.1

© 2019 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the [AMS Copyright Policy](#) (www.ametsoc.org/PUBSReuseLicenses).

to NOAA's 2015–17 Hazardous Weather Testbed (HWT) Spring Forecast Experiments (e.g., Gallo et al. 2017; Clark et al. 2018) and in support of the Plains Elevated Convection at Night (PECAN) field project (Geerts et al. 2017), the National Center for Atmospheric Research (NCAR) produced global, real-time, deterministic, 5-day forecasts with the atmospheric component of the Model for Prediction Across Scales (MPAS; Skamarock et al. 2012) using a variable-resolution mesh featuring 3-km cell spacing over the conterminous United States (CONUS). These MPAS demonstrations yielded credible precipitation and severe weather forecasts through 5 days (e.g., Wong and Skamarock 2016; Gallo et al. 2017).

Yet, 5-day deterministic forecasts, regardless of their resolutions, have limited accuracy due to error growth emanating from imperfect model formulations and initial conditions (e.g., Lorenz 1963). In fact, high-resolution forecast accuracy degrades exceptionally quickly because small-scale errors grow faster than larger-scale errors, especially in areas of moist convection (e.g., Lorenz 1969; Zhang et al. 2003; Hohenegger and Schär 2007). Accordingly, while forecast uncertainty should be considered at all temporal and spatial scales (e.g., Murphy 1993), it is particularly imperative to explicitly quantify uncertainty of high-resolution forecasts, which can be achieved through convection-allowing ensembles (CAEs) that provide probabilistic products. It is now well-recognized that CAEs are needed even for short, hour-long forecasts (e.g., Stensrud et al. 2009; Wheatley et al. 2015; Yussouf et al. 2015), and similarly, longer-term convection-allowing forecasts, including those provided by global NWP models, should be presented probabilistically through ensembles.

While CAEs provide useful guidance (e.g., Clark et al. 2012; Evans et al. 2014; Schwartz et al. 2015a,b, 2019; Sobash et al. 2016a,b), have become operational or semioperational at many worldwide NWP centers (Gebhardt et al. 2011; Peralta et al. 2012; Schwartz et al. 2015b; Hagelin et al. 2017; Raynaud and Bouttier 2017; Klasa et al. 2018), and produce higher-quality short-term forecasts than convection-parameterizing ensembles (Clark et al. 2009; Duc et al. 2013; Iyer et al. 2016; Schellander-Gorgas et al. 2017), CAE forecasts from global models have not received attention. Furthermore, it is unclear whether medium-range probabilistic CAE forecasts can improve upon corresponding coarser-resolution probabilistic forecasts given potential for small-scale errors to amplify with time, eventually sullyng CAE forecasts (e.g., Lorenz

1969; Zhang et al. 2003; Hohenegger and Schär 2007; Judt 2018).

Thus, to address this uncertainty while looking ahead to a future where global convection-allowing NWP models are commonplace, this study examined 3–5-day CAE forecast performance with a 10-member, MPAS-based ensemble employing a variable-resolution global mesh with 3-km cell spacing over the CONUS. The variable-resolution ensemble forecasts were compared to forecasts from both a 10-member MPAS ensemble with global quasi-uniform 15-km cell spacing and NCEP's operational Global Ensemble Forecast System (GEFS). Forecasts were verified with a focus on precipitation, which reflects many physical processes and is regularly used to summarize CAE versus convection-parameterizing ensemble performance (e.g., Clark et al. 2009; Duc et al. 2013; Iyer et al. 2016; Schellander-Gorgas et al. 2017).

The next section describes model configurations and the experimental design, while section 3 details the precipitation verification strategy. Objective performance statistics are presented in sections 4 and 5, results are discussed in section 6, and a case study manifesting aspects of the objective statistics is provided in section 7 before concluding in section 8.

2. Model configurations and experimental design

a. MPAS configurations

All weather forecasts were produced by version 5.1 of MPAS, a nonhydrostatic NWP model with an unstructured centroidal Voronoi mesh that enables variable-resolution configurations with smooth transitions between regions of different resolutions (Skamarock et al. 2012). Two sets of 132-h (5.5-day) 10-member MPAS ensemble forecasts were generated. The first 10-member ensemble had globally quasi-uniform 15-km spacing between cell centers ("MPAS-15km") while the second had a variable-resolution mesh (Fig. 1) with 3-km spacing between cell centers over the CONUS that relaxed to 15 km over the rest of the globe ("MPAS-3km"). The variable-resolution mesh was identical to that used during the successful HWT and PECAN demonstrations.

Every member in both MPAS ensembles had 55 height-based vertical levels (Klemp 2011), a 30-km model top, and identical physical parameterizations (Table 1) that were harvested from version 3.9 of the Advanced Research Weather Research and Forecasting (WRF) Model (WRFv3.9; Skamarock et al. 2008; Powers et al. 2017). The only differences between the quasi-uniform 15-km and variable-resolution 15-/3-km MPAS configurations regarded time step and horizontal

FIG. 1. Approximate distance between cell centers (km) of the variable-resolution global MPAS mesh. Portions of the globe not shown had approximately 15-km spacing between cell centers. The stippled area denotes the verification region (CONUS east of 105°W). Yellow and cyan dots denote locations of rawinsonde observations over the central and eastern CONUS regions, respectively; the division between the central and eastern regions was 87°W, which is annotated.

diffusion. Specifically, ensemble forecasts produced on the quasi-uniform 15-km mesh had a 75-s time step, while the variable-resolution 15-/3-km forecasts had an 18-s time step; tailoring time step to the highest-resolution portion of a variable-resolution mesh is necessary to maintain stability and consistent with previous studies employing variable-resolution MPAS configurations (e.g., [Davis et al. 2016](#); [Fowler et al. 2016](#); [Ha et al. 2017](#)). Explicit horizontal diffusion operators also differed and scaled with the finest cell

spacing in the mesh (e.g., [Skamarock et al. 2014](#); [Fowler et al. 2016](#)). These discrepancies regarding time step and diffusion are direct reflections of the different meshes and do not detract from a robust comparison of sensitivity to horizontal resolution.

Importantly, the “scale aware” Grell–Freitas (G–F) cumulus parameterization ([Grell and Freitas 2014](#); [Fowler et al. 2016](#)) as implemented in WRFv3.9 was used in both 10-member MPAS ensembles. As spacing between cell centers decreased to convection-allowing

TABLE 1. Physical parameterizations used by all members in the MPAS ensembles. These physics schemes were based on implementations in version 3.9 of the WRF Model.

Parameterization	MPAS option	References
Microphysics	Thompson	Thompson et al. (2008)
Longwave and shortwave radiation	Rapid Radiative Transfer Model for global climate models (RRTMG)	Mlawer et al. (1997) ; Iacono et al. (2008)
Planetary boundary layer	Mellor–Yamada–Nakanishi–Niino (MYNN)	Nakanishi and Niino (2009)
Land surface model	Noah	Chen and Dudhia (2001)
Cumulus	Grell–Freitas scale aware	Grell and Freitas (2014)

resolutions, the G–F parameterization effectively became a shallow convection scheme (Fowler et al. 2016). Thus, most precipitation over the 3-km region of the 15-/3-km variable-resolution mesh was explicitly produced, similar to Judt (2018), who noted the G–F scheme yielded little parameterized precipitation in 4-km global MPAS simulations. Conversely, the G–F scheme substantially contributed to precipitation over 15-km areas of the variable-resolution 15-/3-km mesh and the entirety of the quasi-uniform 15-km mesh.

b. Model initialization

Both sets of 132-h MPAS ensemble forecasts were initialized by interpolating 0.5° analyses from 10 perturbation members of NCEP's operational GEFS (e.g., Zhou et al. 2017) onto the global MPAS meshes. The 10 GEFS members differed and provided initial condition diversity—the sole source of diversity—for the MPAS ensemble forecasts. As detailed by Zhou et al. (2017), GEFS initial conditions were produced by adding 6-h forecast perturbations derived from an ensemble Kalman filter data assimilation system (Whitaker and Hamill 2002) to deterministic “hybrid” variational-ensemble analyses produced for NCEP's deterministic Global Forecast System (e.g., Wang and Lei 2014; Kleist and Ide 2015a,b).

Although the GEFS comprised 21 members (20 perturbations plus a control), MPAS forecasts were only initialized from 10 randomly selected perturbation members due to computational considerations; random selection was appropriate because GEFS perturbation members were equally likely (a desirable property when evaluating probabilistic forecasts; e.g., Schwartz et al. 2019). While the necessary number of ensemble members for global CAE forecasts is unknown, limited-area CAEs with 10 members are sufficient to provide skillful and useful guidance (e.g., Clark et al. 2011, 2018; Schwartz et al. 2014).

Because of the relatively coarse-resolution GEFS analyses compared to the MPAS mesh cell spacings, a “spinup” period was expected once MPAS integration began. Results suggested the spinup lasted for ~ 12 h, consistent with previous studies employing MPAS configurations containing a 3-km region (e.g., Wong and Skamarock 2016). Thus, the first 12 forecast hours were excluded from analysis.

c. Reference ensemble

In addition to evaluating performance of the two MPAS ensembles, GEFS forecasts themselves (e.g., Zhou et al. 2017) were verified. Although the GEFS had substantially coarser horizontal resolution (output available at 3-h intervals on a 0.5° grid) and vastly

different physics and dynamics than the MPAS forecasts, GEFS forecasts represented a good baseline to assess whether the higher-resolution, experimental MPAS ensembles were competitive with an operational ensemble. Unlike the MPAS ensembles, the GEFS applied stochastic physics to generate diversity (Hou et al. 2006; Zhou et al. 2016).

d. Experimental period

The two sets of 132-h, 10-member MPAS ensemble forecasts were initialized at 0000 UTC each day between 23 April and 27 May 2017 (inclusive; 35 total cases). GEFS forecasts initialized at 0000 UTC over this period were also examined. During these 5 weeks, several heavy rain and severe weather episodes occurred over the CONUS east of the Rockies (Figs. 2a,b), where the mean 500-hPa flow was relatively strong and generally characterized by broad troughing (Fig. 2c). Most precipitation events over the central and eastern CONUS were strongly forced and occurred ahead of deep upper-level troughs.

3. Precipitation verification strategy

Precipitation forecasts were compared to gridded Stage IV (ST4) data (Lin and Mitchell 2005) from NCEP considered as “truth.” Objective verification was performed over the CONUS east of 105°W (the “verification region”; Fig. 1), where ST4 data were most robust (e.g., Nelson et al. 2016). Analysis focused on 3-h accumulated precipitation, the shortest accumulation interval available from GEFS output.

When possible, verification statistics were computed on native grids. However, many objective verification metrics require that forecast and observed fields share a common grid. Given large resolution differences between the various forecasts, it was important to carefully consider the grid for verification, because verification scores can be sensitive to the verifying grid's resolution (e.g., Gallus 2002; Yan and Gallus 2016).

Thus, to assess sensitivity to the verification grid, all precipitation forecasts were interpolated to the ST4 grid (4.763-km horizontal grid spacing) and common 15- and 40-km grids. As expected, verification scores were best on the 40-km grid (e.g., Gallus 2002), but identical qualitative conclusions were obtained on all three grids. Ultimately, only results for verification on the ST4 grid are presented, following Wolff et al. (2014) and Mittermaier (2018), who advised interpolating model forecasts to a finescale observation grid when the question of interest is determining whether a model can reproduce finescale events and the finescale observations can be trusted as physically

FIG. 2. (a) Total accumulated Stage IV (ST4) precipitation (mm) between 0000 UTC 23 Apr and 1200 UTC 1 Jun 2017 east of 105°W, coinciding with the verification region (Fig. 1). Gray shaded points were either outside the range of the ST4 observations or had missing data in at least one 3-h accumulation interval. (b) Severe weather reports received by the Storm Prediction Center east of 105°W between 0000 UTC 23 Apr and 1200 UTC 1 Jun 2017, where blue, green, and red markers indicate severe wind, severe hail, and tornado reports, respectively. The dashed line is 87°W, which separates the eastern and central regions discussed in section 6. (c) Average 500-hPa wind speed (shaded; kts) and height (m; contoured every 40 m) over all 0000 and 1200 UTC Global Forecast System analyses between 0000 UTC 23 Apr and 1200 UTC 1 Jun 2017 (inclusive). The accumulation/averaging period in (a)–(c) encompasses all possible valid times of the model forecasts.

realistic. Interpolating to the ST4 grid meant a slight upscaling of the 3-km forecasts, while interpolating GEFS and 15-km MPAS forecasts to the finer ST4 grid did not introduce additional detail and can be viewed as projecting coarser output onto a finer grid (Mittermaier 2018).

GEFS precipitation forecasts were interpolated to the ST4 grid with a budget interpolation algorithm that conserves total precipitation (e.g., Accadia et al. 2003), the typically preferred method for re-gridding precipitation forecasts. However, application of a conservative algorithm (e.g., Jones 1999) to interpolate from the unstructured MPAS mesh to the ST4 grid sometimes yielded interpolated fields that locally did not visually resemble native fields. Thus, barycentric interpolation (e.g., Skamarock et al. 2014; Ha et al. 2017), resembling bilinear interpolation, was used to interpolate MPAS precipitation forecasts to the ST4 grid. This choice did not meaningfully impact MPAS precipitation distributions throughout the diurnal cycle (Fig. 3), increasing confidence that barycentric interpolation had minimal impacts on verification statistics.

4. Precipitation climatologies

a. Areal coverages

Aggregate areal coverages of 3-h accumulated precipitation meeting or exceeding selected physical accumulation thresholds [e.g., 10.0 mm (3 h)^{−1}] were

calculated over the verification region on native grids to assess how the models replicated the observed diurnal cycle and precipitation distributions (Fig. 4). Regarding diurnal cycle representation, all ensembles usually peaked early compared to observations for thresholds ≤ 2.5 mm (3 h)^{−1} (Figs. 4a,b). For thresholds between 5.0 and 25.0 mm (3 h)^{−1} (Figs. 4c–e), 15-km MPAS forecasts regularly peaked 6 h too late, with the GEFS and 3-km MPAS peaks typically within 3 h of those observed. At the 50.0 mm (3 h)^{−1} threshold, both MPAS ensembles correctly represented the diurnal cycle, while a clear diurnal cycle vanished in the GEFS (Fig. 4f).

In terms of magnitude, 3-km MPAS forecasts usually had peak coverages closest to those observed (Fig. 4). The GEFS overpredicted for 3-h precipitation accumulations ≤ 2.5 mm but underpredicted at higher thresholds, with worsening low biases as thresholds increased; GEFS resolution was simply too coarse to capture intense convective rainfall. Conversely, 15-km MPAS ensemble members noticeably overpredicted at all thresholds.

Further analysis indicated the 15-km MPAS ensemble's extreme overprediction at light rainfall rates (Figs. 4a,b) was due to an overactive cumulus parameterization, as areal coverages of precipitation produced by the convection scheme (as opposed to explicitly produced precipitation) were substantially larger in the 15-km MPAS ensemble than those from the GEFS and regularly exceeded the total observed

FIG. 3. Probability density functions (PDFs; %) aggregated over all 35 (a) 18-, (b) 24-, (c) 30-, and (d) 36-h forecasts of 3-h accumulated precipitation (mm) for ensemble member 1 on the native (solid) and ST4 (dashed) grids over the verification region (CONUS east of 105°W) from the 3-km MPAS ensemble (MPAS-3km) and 15-km MPAS ensemble (MPAS-15km). Similar results held for other members. The highest values of the leftmost points are approximately 95%.

precipitation (not shown). Moreover, this gross overprediction of light precipitation could be dramatically reduced by modifying the cumulus parameterization, as evidenced by forecasts from a third, auxiliary, MPAS ensemble identical to the 15-km quasi-uniform ensemble except for a very different version of the G–F cumulus parameterization as implemented in version 3.8 of the WRF Model. However, these improvements regarding precipitation climatology did not translate into better spatial skill, which was considerably degraded in this auxiliary ensemble compared to the primary 15-km ensemble described in section 2a. This sensitivity of the 15-km MPAS forecasts to different G–F implementations underscores considerable challenges of tuning convective parameterization schemes.

Ultimately, results from the auxiliary 15-km ensemble are not further discussed for two reasons:

- 1) When determining whether the variable-resolution ensemble (Fig. 1) had better spatial skill than a quasi-uniform 15-km ensemble—a main goal of this work—the 15-km ensemble using the G–F cumulus parameterization in WRFv3.9 represented a higher standard, despite poor biases at light precipitation rates.
- 2) The variable-resolution ensemble used the G–F scheme from WRFv3.9, meaning the 15-km MPAS ensemble also using WRFv3.9’s cumulus parameterization provided the fairest comparison.

b. Domain-total precipitation

Compensating GEFS biases at high and low precipitation rates (Fig. 4) yielded total accumulated precipitation over the verification region closely matching ST4 observations (Fig. 5). The 3-km MPAS forecasts also produced total precipitation similar to that observed, and consistent with overprediction at all thresholds, the 15-km

FIG. 4. Fractional areal coverage (%) of 3-h accumulated precipitation meeting or exceeding (a) 1.0, (b) 2.5, (c) 5.0, (d) 10.0, (e) 25.0, and (f) 50.0 mm (3 h)⁻¹ over the verification region (CONUS east of 105°W), computed on native grids, and aggregated over all 35 forecasts as a function of forecast hour. The orange, purple, and green shadings represent envelopes of the 10 members comprising the 3-km MPAS ensemble (MPAS-3km), 15-km MPAS ensemble (MPAS-15km), and GEFS, respectively, and darker shadings indicate intersections of two (or more) ensemble envelopes. Values on the x axis represent ending forecast hours of 3-h accumulation periods (e.g., an x -axis value of 24 is for 3-h accumulated precipitation between 21 and 24 h). Dashed vertical lines indicate times when the ST4 observations (black curves) reached their diurnal maxima.

MPAS forecasts produced too much total precipitation. Both the GEFS and 3-km MPAS ensemble featured clear diurnal cycles that corresponded well with observations (Fig. 5), although 3-km forecasts usually peaked 3 h early. Conversely, the 15-km MPAS ensemble poorly represented the diurnal cycle of total precipitation, largely due to its acute overprediction at light precipitation rates. In fact, by assigning values of zero (i.e., no precipitation) to those 15-km MPAS cells with precipitation < 2.5 mm (3 h)⁻¹, double peaks were removed, yielding single-peaked diurnal patterns similar to the other ensembles (not shown).

c. Summary

Overall, precipitation climatologies clearly revealed benefits of 3-km cell spacing throughout the 132-h forecasts (Figs. 4, 5). Although the 3-km MPAS ensemble had small errors regarding timing of diurnal maxima, its areal

coverage magnitudes were closest to those observed across all thresholds. These findings are consistent with previous studies showing convection-allowing models typically represent the diurnal cycle better than convection-parameterizing models (e.g., Clark et al. 2007, 2009; Weisman et al. 2008; Schwartz et al. 2009; Iyer et al. 2016; Schellander-Gorgas et al. 2017) and demonstrate these benefits can extend through 5-day forecasts.

5. Evaluation of probabilistic precipitation forecasts and spread

a. Methods

1) ADDRESSING BIAS

Areal coverages indicated obvious biases, particularly in the GEFS and 15-km MPAS ensemble. Unfortunately,

FIG. 5. Average 3-h accumulated precipitation (mm) per grid point over all 35 forecasts and the verification region (CONUS east of 105°W) computed on native grids as a function of forecast hour. The orange, purple, and green shadings represent envelopes of the 10 members comprising the 3-km MPAS ensemble (MPAS-3km), 15-km MPAS ensemble (MPAS-15km), and GEFS, respectively, and darker shadings indicate intersections of two ensemble envelopes. Values on the x axis represent ending forecast hours of 3-h accumulation periods (e.g., an x -axis value of 24 is for 3-h accumulated precipitation between 21 and 24 h). Dashed vertical lines indicate times when the ST4 observations (black curves) reached their diurnal maxima.

large biases muddle interpretation of verification metrics designed to quantify spatial displacement errors (e.g., Baldwin and Kain 2006; Roberts and Lean 2008), and using physical accumulation thresholds to define events (as in Fig. 4) preserves these complicating biases.

Thus, for probabilistic verification, percentile thresholds (e.g., the 95th percentile, which selects the top 5% of values) were used to define events, which controls for bias and permits a robust assessment of spatial performance within the context of a model's climate (e.g., Roberts and Lean 2008; Mittermaier and Roberts 2010; Mittermaier et al. 2013; Dey et al. 2014; Gowan et al. 2018; Woodhams et al. 2018). Practically, selecting a percentile threshold establishes corresponding physical thresholds ultimately used when querying forecast and observed fields to determine event occurrence. Although the forecast and observed physical thresholds differ for a fixed percentile threshold if a forecast is biased (e.g., the top 5% of forecast and observed 3-h precipitation accumulations may exceed 10.0 and 20.0 mm, respectively), because the forecast and observations exceed their respective physical thresholds at an equal number of points, bias is removed.

Percentiles ranging from 90% to 99.9% were chosen to encompass both lighter, broader precipitation features as well as intense, localized convective elements. Physical thresholds corresponding to fixed percentile thresholds were calculated separately for each ensemble member and observations at each forecast output time on the ST4 grid. Only grid points within the verification region (Fig. 1) participated in determining mappings between percentile and physical thresholds. If the physical threshold corresponding to the selected percentile threshold was zero (i.e., no precipitation) for a particular 3-h accumulation interval in either the forecast or observations, that accumulation interval was excluded from aggregate statistics at the selected percentile threshold. However, these situations were rare.

Mean physical thresholds (Fig. 6) corresponding to selected percentile thresholds revealed similar patterns and biases as areal coverages (Fig. 4); indeed, areal coverage exceedances of physical thresholds and percentile distributions provide complementary information. As the percentile threshold increased, differences between the corresponding physical thresholds for the three ensembles also increased, and at the 97th percentile and above, signals of the large 15-km biases at light precipitation rates were gone.

2) PROBABILITY GENERATION

After interpolating precipitation forecasts to the ST4 grid (i.e., section 3), several steps were performed to produce probabilistic fields suitable for verification following Schwartz and Sobash (2017). First, a binary probability (BP) at the i th point for the j th ensemble member (BP_{ij}) was computed as

$$BP_{ij} = \begin{cases} 1 & \text{if } f_{ij} \geq q_j \\ 0 & \text{otherwise} \end{cases}, \quad (1)$$

where f_{ij} is the forecast precipitation at the i th point for the j th member and q_j is the physical event threshold for the j th member. Equation (1) applies to both physical and percentile event definitions: if physical thresholds (e.g., Fig. 4) are chosen, q_j is identical for all ensemble members, but if percentile thresholds are used, the physical threshold corresponding to the selected percentile threshold differs for each member (e.g., Fig. 6) leading to unique q_j .

After obtaining BPs, gridscale ensemble probabilities (EPs) at the i th point (EP_i) were determined simply as

$$EP_i = \frac{1}{N} \sum_{j=1}^N BP_{ij}, \quad (2)$$

where N is the ensemble size (i.e., $N = 10$ for the three ensembles). EPs were then used to derive probabilities

FIG. 6. Average physical thresholds [mm (3 h)⁻¹] over all 35 forecasts corresponding to the (a) 90th, (b) 95th, (c) 97th, (d) 99th, (e) 99.5th, and (f) 99.9th percentile thresholds as a function of forecast hour. The physical thresholds were computed separately for each day and 3-h forecast period on the ST4 grid over the verification region (CONUS east of 105°W) and averaged to obtain the y-axis values. The orange, purple, and green shadings represent envelopes of the 10 members comprising the 3-km MPAS ensemble (MPAS-3km), 15-km MPAS ensemble (MPAS-15km), and GEFS, respectively, and darker shadings indicate intersections of two (or more) ensemble envelopes. Values on the x axis represent ending forecast hours of 3-h accumulation periods (e.g., an x-axis value of 24 is for 3-h accumulated precipitation between 21 and 24 h). Dashed vertical lines indicate times when the ST4 observations (black curves) reached their diurnal maxima.

reflecting a “neighborhood” approach (e.g., Theis et al. 2005; Ebert 2008, 2009), which addresses spatial uncertainties and recognizes the inherent inaccuracy of high-resolution NWP models at individual grid points. Specifically, “neighborhood ensemble probabilities” (NEPs; Schwartz et al. 2010; Schwartz and Sobash 2017) were obtained by first selecting a neighborhood length scale (r) that determines how many points reside within the neighborhood of a given point. Then, letting S_i denote the unique set of N_b points within the neighborhood of the i th point, the NEP at the i th point (NEP _{i}) was computed by averaging EPs over all grid points within the neighborhood of the i th point:

$$\text{NEP}_i = \frac{1}{N_b} \sum_{k=1}^{N_b} \text{EP}_k, \quad k \in S_i. \quad (3)$$

NEPs are interpreted as probabilities of event occurrence at the grid scale given a neighborhood length scale (Schwartz and Sobash 2017) and were produced from the three ensembles for r between 5 and 400 km, which represented radii of circular neighborhoods. Following Schwartz and Sobash (2017), NEPs at the i th point were objectively verified against corresponding observations (i.e., ST4) at the i th point, where, depending on the verification metric, the i th observed value could either be binary (i.e., 0 or 1) or fractional; fractional observations (e.g., Roberts and Lean 2008) at the i th point were obtained by applying Eqs. (1)–(3) to ST4 grids with $N = 1$.

3) STATISTICAL SIGNIFICANCE TESTING

Statistical significance of various metrics was determined with a bootstrap technique by randomly drawing

paired samples (10 000 times) of daily statistics from two ensembles over all forecast cases to calculate resampled distributions of aggregate differences between two ensembles (e.g., Hamill 1999; Wolff et al. 2014; Schwartz et al. 2015a). This procedure assumed individual forecasts, initialized 24 h apart, were independent (e.g., Hamill 1999). Bounds of 90% bootstrap confidence intervals (CIs) were obtained from the distribution of resampled aggregate differences using the bias corrected and accelerated method (e.g., Gilleland 2010). If bounds of a 90% bootstrap CI did not encompass zero, using a one-tailed interpretation, differences between two ensembles were statistically significant at the 95% level or higher.

b. Results

1) ROC AREAS

Areas under the relative operating characteristic (ROC) curve (Mason 1982; Mason and Graham 2002) assessed the ensembles' abilities to discriminate events from climatology and were computed with a trapezoidal approximation using probabilistic thresholds of 1%, 2%, 3%, 4%, 5%, 10%, 15%, ..., 95%, and 100%. An ensemble has successful discriminating ability compared to random forecasts if ROC areas are >0.5 .

ROC areas for NEPs computed with $r = 100$ km remained above 0.5 for all thresholds throughout the 132-h forecast but decreased with forecast length, indicating forecast degradation with time (Fig. 7). Diurnal variations were evident in all three ensembles and especially pronounced for the 15-km MPAS ensemble at the 99th percentile and below. Given that percentile thresholds were employed, these diurnal variations were not attributable to the sometimes large 15-km MPAS biases (e.g., Figs. 4a,b). Rather, the oscillations suggest the 15-km MPAS ensemble performed relatively poorly during convective initiation and comparatively better during upscale evolution when systems increased in size and became more intense, as manifested by correspondence between maximum 15-km MPAS ensemble ROC areas and physical event thresholds (Fig. 6).

Through 48 h at the 90th, 95th, and 97th percentile thresholds, the 3-km MPAS ensemble consistently had statistically significantly higher ROC areas than the other ensembles, while GEFS ROC areas were statistically significantly lowest (Figs. 7a–c). However, between 48 and 72 h, while GEFS ROC areas remained statistically significantly lowest, the 3- and 15-km MPAS ensembles had more similar ROC areas. After 72 h, there were fewer statistically significant differences between the three ensembles than at earlier

forecast lengths, although the GEFS continued to have the lowest ROC areas.

At higher percentile thresholds, the two MPAS ensembles usually had statistically significantly higher ROC areas than the GEFS throughout the 132-h forecast (Figs. 7d–f). As at lower thresholds, the 3-km MPAS ensemble was typically statistically significantly better than the 15-km MPAS ensemble over the first 48 h, with only occasional statistically significant differences thereafter. ROC areas computed for NEPs generated with other neighborhood length scales revealed similar qualitative differences between the three ensembles as the 100-km results (not shown).

2) FRACTIONS SKILL SCORES

The popular fractions skill score [FSS; Roberts and Lean (2008)] was also used to assess spatial placement. FSSs of 1 indicate perfect forecasts while $FSS = 0$ indicates no skill. In addition, Roberts and Lean (2008) showed how the neighborhood length scale yielding $FSS = FSS_{\text{useful}}$ corresponds to the minimum “useful” neighborhood length scale, where $FSS_{\text{useful}} = 0.5 + f_0/2$ and f_0 is the fraction of observed events over the verification region. For verification based on percentile thresholds, f_0 is specified by the chosen threshold (e.g., the 95th percentile threshold selects the top 5% of events, so $f_0 = 0.05$).

Like ROC areas, FSSs for NEPs computed with $r = 100$ km revealed diurnal signals, especially for the 15-km MPAS ensemble at lower thresholds (Fig. 8). Otherwise, FSSs decreased with forecast length for all ensembles and dropped below FSS_{useful} progressively earlier as the threshold increased, reflecting greater difficulty of predicting localized features than widespread events.

Through ~ 48 h at the 97th percentile threshold and below, the 3-km MPAS ensemble had the highest FSSs and the GEFS the lowest (Figs. 8a–c). Thereafter, FSSs for all ensembles were typically comparable. At higher thresholds (Figs. 8d–f), both MPAS ensembles usually had higher FSSs than the GEFS throughout the forecast, but differences with respect to the GEFS were not always statistically significant. As at lower thresholds, for the first 48 h the 3-km MPAS ensemble had statistically significantly higher FSSs than the 15-km MPAS ensemble, but differences between the two MPAS ensembles were small after ~ 48 h (Figs. 8d–f).

Minimum useful neighborhood length scales for the 3-km MPAS ensemble steadily increased with forecast length and percentile threshold, indicating poorer forecasts for longer lead times and more extreme events (Fig. 9). For example, at the 90th percentile, day 1 (18–36-h) forecasts had a useful neighborhood length scale of 5 km, but day 5 (114–132 h) forecasts

FIG. 7. Areas under the ROC curve over the verification region (CONUS east of 105°W) with a 100-km neighborhood length scale for the (a) 90th, (b) 95th, (c) 97th, (d) 99th, (e) 99.5th, and (f) 99.9th percentile thresholds aggregated over all 35 forecasts as a function of forecast hour. Values on the x axis represent ending forecast hours of 3-h accumulation periods (e.g., an x-axis value of 24 is for 3-h accumulated precipitation between 21 and 24 h). Symbols along the top axis denote forecast hours when differences between two ensembles were statistically significant at the 95% level. The top row indicates differences between the 3- and 15-km MPAS ensembles (MPAS-3km vs MPAS-15km), the middle row shows MPAS-3km vs GEFS, and the bottom row is for MPAS-15km vs GEFS. Colors of the symbols correspond to the legend and denote the ensemble with the statistically significantly higher ROC area. For example, in the top row, orange symbols indicate MPAS-3km had statistically significantly higher ROC areas than MPAS-15km and purple symbols mean MPAS-15km had statistically significantly higher ROC areas than MPAS-3km. Absence of a symbol means the differences were not statistically significant at the 95% level.

only achieved the minimum useful scale between 100 and 150 km (Fig. 9a), and for the most extreme events, the useful scale increased from ~150 km on day 1 to ~400 km by day 5 (Fig. 9f). These results suggest model output should be presented to forecasters with increasing neighborhood length scales (i.e., increased smoothing) as forecasts progress. Furthermore, differences between days 2 and 3, where FSSs dropped particularly sharply at the 97th percentile threshold and above (Figs. 9c–f), may require special attention.

3) ATTRIBUTES STATISTICS

Attributes diagrams (Wilks 2011) were constructed using forecast probability bins of 0%–5%, 5%–15%, 15%–25%, ..., 85%–95%, and 95%–100% (Figs. 10, 11).

Perfect reliability was achieved for curves on the diagonal, points within shaded areas had skill compared to forecasts of climatology, and values were not plotted for bins with <500 samples.

At the 90th, 95th, and 97th percentile thresholds for NEPs computed with $r = 100$ km, the 3-km MPAS ensemble had the best—near-perfect—reliability for day 1 forecasts and GEFS reliabilities were typically worst (Figs. 10a–c). Broadly similar behaviors occurred for day 2 (42–60-h) forecasts, although 3-km MPAS reliabilities were further from perfect and there were fewer instances of statistically significant differences between the 15- and 3-km MPAS ensembles (not shown). By day 3 (66–84-h forecasts), there were mostly insignificant differences between the two

FIG. 8. As in Fig. 7, but for aggregate FSSs. Values of FSS_{useful} are shown by dashed horizontal lines.

MPAS ensembles while the GEFS continued to have the worst reliabilities (Figs. 10d–f). The three ensembles generally became even more similar by day 5 (Figs. 10g–i), although some statistically significant differences remained. At these lower thresholds, all three ensembles had skill compared to forecasts of climatology for most probability bins.

For higher thresholds (Fig. 11), day 1 and 2 reliabilities generally mirrored those at lower thresholds (e.g., Fig. 10), with the 3-km ensemble providing the best reliability (Figs. 11a–c). By days 3–5, the 3- and 15-km MPAS ensembles had comparable reliabilities that were usually better than GEFS reliabilities, although differences with respect to the GEFS were not always statistically significant (Figs. 11d–i). At these higher thresholds, both MPAS ensembles were typically skillful with respect to climatological forecasts while the GEFS often was not, and as lead time increased, the ensembles became progressively incapable of producing higher NEPs (Figs. 10, 11), reflecting increasing spread.

Changing r can impact attributes statistics by modifying the amount of smoothing and resulting probability

distributions (e.g., Schwartz and Sobash 2017). For these forecasts, reliability was closest to perfect for most percentile thresholds using 100- and 150-km neighborhood length scales, and attributes statistics computed with other neighborhood length scales were usually overconfident and underconfident for $r < 100$ km and $r > 150$ km, respectively.

Overall, while the 3-km MPAS ensemble had excellent day 1 reliability for $r = 100$ km, as forecasts progressed, all ensembles' reliabilities degraded and became increasingly overconfident. By days 3–5, there were only occasionally statistically significant differences between the 3- and 15-km MPAS ensembles, and these results suggest achieving perfect reliability in 3–5-day variable-resolution CAE forecasts with a global model may be challenging.

4) ENSEMBLE VARIANCE

While attributes diagrams provided some indications of ensemble spread characteristics, spread was further assessed by examining gridpoint variance of precipitation over the verification region. Given the

FIG. 9. FSSs as a function of neighborhood length scale (km) computed over the verification region (CONUS east of 105°W) for the 3-km MPAS ensemble's forecasts of 3-h accumulated precipitation, aggregated over all 35 forecasts and various forecast ranges (legend) for the (a) 90th, (b) 95th, (c) 97th, (d) 99th, (e) 99.5th, and (f) 99.9th percentile thresholds. Values of FSS_{useful} are shown by dashed horizontal lines.

sometimes large biases between ensembles (e.g., Fig. 4), to equitably compare variances across the ensembles, variances were computed for 3-h accumulated precipitation fields that were bias corrected with a probability matching technique that forced each ensemble member's distribution to the ST4 distribution by replacing the model *grid point* containing the highest precipitation with the highest ST4 *amount*, and so on (e.g., Ebert 2001; Clark et al. 2009, 2010a,b; Schwartz et al. 2015a). This approach preserves spatial patterns within each member, and, similar to using percentile thresholds, permits an evaluation of performance without impacts of bias.

Variances based on bias corrected fields were largest in the 3-km MPAS ensemble and increased with forecast length (Fig. 12a), similar to Clark et al. (2010b), who found a CAE had more precipitation variance than a

coarser-resolution ensemble in multiphysics ensemble frameworks. To assess variance growth, the variance growth rate (VGR) was defined after Clark et al. (2010b) as

$$\text{VGR} = \left(\frac{\text{Var}_f - \text{Var}_{15}}{\text{Var}_{15}} \right) \times 100\%, \quad (4)$$

where Var_{15} is the reference variance at forecast hour 15 and Var_f is the variance at forecast hour f for a specific ensemble; hour 15 was chosen as the reference lead time because it was after the spinup period, yet relatively early in the forecast.

Equation (4) indicates VGRs are sensitive to the reference value (i.e., Var_{15}), which differed across the ensembles (Fig. 12a). Therefore, it is unsurprising that the GEFS, which had the smallest Var_{15} , had the highest

FIG. 10. Attributes diagrams computed over the verification region (CONUS east of 105°W) with a 100-km neighborhood length scale aggregated over all 35 (a)–(c) 18–36-, (d)–(f) 66–84-, and (g)–(i) 114–132-h forecasts for the (a),(d),(g) 90th, (b),(e),(h) 95th, and (c),(f),(i) 97th percentile thresholds. Horizontal lines near the x axis represent observed frequencies of the event, diagonal lines are lines of perfect reliability, and open circles show forecast frequencies (%) within each probability bin, with colors corresponding to the legend. Points lying in gray-shaded regions had skill compared to climatological forecasts as measured by the Brier skill score (Brier 1950; Wilks 2011). Symbols along the top axis denote forecast hours when differences between two ensembles were statistically significant at the 95% level. The top row indicates differences between the 3- and 15-km MPAS ensembles (MPAS-3km vs MPAS-15km), the middle row shows MPAS-3km vs GEFS, and the bottom row is for MPAS-15km vs GEFS. Colors of the symbols correspond to the legend and denote the ensemble with statistically significantly better reliability (i.e., the curve closest to perfect reliability). For example, in the top row, orange symbols indicate MPAS-3km had statistically significantly better reliability than MPAS-15km and purple symbols mean MPAS-15km had statistically significantly better reliability than MPAS-3km. Absence of a symbol means the differences were not statistically significant at the 95% level. Values were not plotted for a particular bin if fewer than 500 grid points had forecast probabilities in that bin over the verification region and all 35 forecasts. Note that the attributes diagrams themselves stop at 100%; area above 100% was added to make room for statistical significance markers.

VGRs throughout the forecast, followed by the two MPAS ensembles, whose VGRs were similar (Fig. 12b). To permit an assessment of VGR relative to a common initial variance, variance was normalized as in Clark et al. (2010b) using

$$\text{Var}_f^* = \text{Var}_f - (\text{Var}_{15} - \text{Var}_{15}^{\text{MPAS-3km}}), \quad (5)$$

where Var_f^* is the adjusted variance at forecast hour f for a specific ensemble and $\text{Var}_{15}^{\text{MPAS-3km}}$ is the variance at forecast hour 15 for the 3-km MPAS ensemble. Thus, Eq. (5) adjusted all ensembles to identical variances at $f = 15$ (i.e., $\text{Var}_{15}^* = \text{Var}_{15}^{\text{MPAS-3km}}$ for all ensembles) and shifted the GEFS and 15-km MPAS VGR time series by a constant.

FIG. 11. As in Fig. 10, but for the (a),(d),(g) 99th, (b),(e),(h) 99.5th, and (c),(f),(i) 99.9th percentile thresholds.

When replacing Var_f by Var_f^* and Var_{15} by Var_{15}^* in Eq. (4), VGRs were more similar (Fig. 12c) compared to unadjusted VGRs (Fig. 12b). Nonetheless, around the first diurnal maxima, 3-km MPAS ensemble VGRs were statistically significantly highest, consistent with rapid upscale growth of small-scale errors (e.g., Lorenz 1969; Zhang et al. 2003; Hohenegger and Schär 2007). Conversely, absence of convective-scale motions in the GEFS and 15-km MPAS ensemble initially limited VGRs. However, as the GEFS and 15-km MPAS forecasts progressed, errors due to cumulus parameterization eventually compensated for missing small-scale errors, and stochastic physics (Hou et al. 2006; Zhou et al. 2016) further contributed to GEFS VGRs. Thus, after the first diurnal cycle, GEFS VGRs were regularly statistically significantly highest while differences between the 15- and 3-km MPAS ensembles' VGRs were usually statistically insignificant (Fig. 12c). Accordingly, by the fifth diurnal cycle, differences

between the variance magnitudes—while still sometimes statistically significant—were at their smallest (Fig. 12a).

6. Discussion

Forecasts at early lead times are strongly influenced by initial conditions, which were provided by coarse 0.5° analyses that lacked storm-scale errors and fostered uncontaminated local environments in which convection could organically develop (e.g., Potvin et al. 2017). These clean, large-scale initial states permitted intrinsic benefits of explicitly represented convection to be realized, and the 3-km MPAS ensemble's superior performance through 48 h is consistent with previous findings that CAEs initialized from substantially coarser analyses produce better precipitation forecasts than convection-parameterizing ensembles through 48 h (e.g., Clark et al. 2009; Duc et al. 2013; Iyer et al. 2016; Schellander-Gorgas et al. 2017).

FIG. 12. Average (a) variance (mm^2), (b) unadjusted variance growth rate (%), and (c) adjusted variance growth rate (%) over the verification region (CONUS east of 105°W) and all 35 forecasts of bias corrected 3-h accumulated precipitation. Variance growth rates are relative to the average variances at forecast hour 15 [first point on the x axis in (a)], and the adjustment in (c) was performed with respect to the 3-km MPAS ensemble (see text). Values on the x axis represent ending forecast hours of 3-h accumulation periods (e.g., an x-axis value of 24 is for 3-h accumulated precipitation between 21 and 24 h). Symbols along the top axis denote forecast hours when differences between two ensembles were statistically significant at the 95% level. The top row indicates differences between the 3- and 15-km MPAS ensembles (MPAS-3km vs MPAS-15km), the middle row shows MPAS-3km vs GEFS, and the bottom row is for MPAS-15km vs GEFS. Colors of the symbols correspond to the legend and denote the ensemble with the statistically significantly higher average values. For example, in the top row, orange symbols indicate MPAS-3km had statistically significantly higher average values than MPAS-15km and purple symbols mean MPAS-15km had statistically significantly higher average values than MPAS-3km. Absence of a symbol means the differences were not statistically significant at the 95% level.

The more similar ROC areas, FSSs, and attributes statistics between 48 and 132 h, particularly between the 15- and 3-km MPAS ensembles, is a novel result with no basis for comparison and hard to definitively explain. It is possible this convergence could have resulted from assigning corresponding members within each ensemble common 0.5° initial conditions that may have posed a large-scale constraint on forecast evolution (e.g., Durran and Gingrich 2014; Durran and Weyn 2016; Potvin et al. 2017; Weyn and Durran 2017, 2019). In addition, diminishing differences with time regarding reliability and spread (Figs. 10–12) could have been facilitated by employing a variable-resolution mesh (Fig. 1), as Nutter et al. (2004) showed how coarse LBCs can limit spread in regional ensembles and suggested similar mechanisms could limit spread in global variable-resolution ensembles because “advection of scale-deficient flow into the high-resolution region will constrain error growth rates at small scales.” It is also possible that small-scale error growth saturated by ~ 48 h in the 3-km ensemble, contributing to comparable spread and reliability as the coarser ensembles thereafter. Furthermore, loss of predictability on increasingly large scales, which impacts all models, could also have led to smaller differences after ~ 48 h. Comprehensively examining these hypotheses is not straightforward in real-data

simulations and beyond the scope of this initial look at global CAEs with a variable-resolution model.

However, another, easily testable, hypothesis regarding the relatively similar 3- and 15-km MPAS ensemble performances after 48 h concerns the possibility that upscale growth of small-scale errors may have eventually contaminated larger scales and hampered precipitation placement in the 3-km forecasts; this hypothesis is consistent with previous studies noting predictability from convection-allowing forecasts over the central-eastern CONUS is lost on scales < 100 km by 48 h (e.g., Zhang et al. 2003; Surcel et al. 2017). Thus, to examine large-scale errors, 15- and 3-km MPAS ensemble mean forecasts were compared to rawinsonde observations over the verification region (Fig. 1). At 850, 700, and 500 hPa, the 3-km ensemble had cold biases over the central CONUS (yellow dots in Fig. 1) that grew with time and were often statistically significantly worse than 15-km biases; RMSEs also typically favored the 15-km ensemble, especially from 60 h onward (Figs. 13a–c). Consistent with low- and midtropospheric cold biases, the 3-km ensemble had low 500-hPa height biases that were statistically significantly worse than those from the 15-km ensemble after 48 h (Fig. 13d). Conversely, over the eastern CONUS (cyan dots in Fig. 1), the 15- and 3-km ensembles had similar 500-hPa height biases (Fig. 13h),

FIG. 13. (a),(e) 850-, (b),(f) 700-, and (c),(g) 500-hPa temperature (K) and (d),(h) 500-hPa height (m) biases (dashed lines; convention is forecast minus observations) and RMSEs (solid lines) for 3- and 15-km MPAS ensemble mean forecasts compared to rawinsonde observations over the (a)–(d) central and (e)–(h) eastern CONUS aggregated over all 35 forecasts as a function of forecast hour. See Fig. 1 for locations of the rawinsonde observations. Open circles denote instances when a particular ensemble had statistically significantly better RMSEs or biases at the 95% level compared to the other ensemble. For example, a purple circle on a dashed line indicates the 15-km MPAS ensemble (MPAS-15km) had a statistically significantly better bias (i.e., closer to zero) than the 3-km MPAS ensemble (MPAS-3km), while an orange circle on a solid line means MPAS-3km had a statistically significantly better (lower) RMSE than MPAS-15km. Absence of a symbol means a difference was not statistically significant at the 95% level.

850-hPa temperature statistics suggested better 3-km performance (Fig. 13e), and although 700- and 500-hPa 15-km errors were often smallest, there were fewer statistically significant differences between the two ensembles than over the central CONUS (Figs. 13f,g). Comparing MPAS output to ERA-Interim reanalyses (e.g., Dee et al. 2011) affirmed these rawinsonde-based results that 3-km large scales were especially poor relative to 15-km large scales over the central CONUS after 48–60 h (not shown). These ERA-Interim-based results, coupled with the fact that identical conclusions were obtained when upscaling 3- and 15-km MPAS forecasts to 0.5° and comparing to rawinsondes, suggests “double-penalty” effects (e.g., Ebert 2008) did not cause the often worse 3-km verification scores.

These large-scale differences between the 3- and 15-km forecasts were due to different representations of convection (i.e., explicit versus parameterized). Moreover, rawinsonde verification, particularly over the central CONUS—where deep convection was common and responsible for two-thirds of all storm reports east of 105°W (Fig. 2b)—supports the hypothesis that accumulation and upscale growth of small-scale errors emanating from explicitly allowed

convection may have substantially degraded 3-km large scales. Although determining the processes responsible for enhanced 3-km errors requires further work, 3-km tropospheric cold biases over the central CONUS may be related to insufficient latent heat release on the convection-allowing portion of the mesh, whereas warming from the G–F cumulus parameterization may have limited 15-km forecast cold biases over the central CONUS.

Accordingly, to the extent large-scale fields modulate precipitation placement in strongly forced situations like those prevailing over the experimental period, these findings suggest bigger 3-km large-scale errors may have eventually counteracted inherent benefits of finer grid spacing, leading to more similar 3- and 15-km precipitation skill and reliability with time. In fact, FSSs and ROC areas indicated statistically significantly better 3-km precipitation forecasts were less frequent over the central CONUS than eastern CONUS (Fig. 14), suggesting benefits of 3-km cell spacing for precipitation forecasting were most diminished when 3-km large-scale errors were routinely worse than 15-km large-scale errors (e.g., Figs. 13a–d). But, when the 15- and 3-km ensembles had comparable large-scale errors, such as for lead times <48 h, intrinsic benefits of explicitly allowed

(a) FSS

	Threshold (%)	Forecast hour																					
		15	18	21	24	27	30	33	36	39	42	45	48	51	54	57	60	63	66	69	72	75	78–132
Eastern CONUS	90.0																						
	95.0																						
	97.0																						
	99.0																						
	99.5																						
	99.9																						93, 117, 132
Central CONUS	90.0																						
	95.0																						
	97.0																						
	99.0																						
	99.5																						
	99.9																						

(b) ROC area

	Threshold (%)	Forecast hour																					
		15	18	21	24	27	30	33	36	39	42	45	48	51	54	57	60	63	66	69	72	75	78–132
Eastern CONUS	90.0																						117, 120
	95.0																						96, 117, 120
	97.0																						93, 96, 117, 120
	99.0																						96, 117, 120
	99.5																						96, 120
	99.9																						
Central CONUS	90.0																						
	95.0																						117
	97.0																						93, 117
	99.0																						
	99.5																						
	99.9																						

FIG. 14. Statistical significance tables for (a) FSSs and (b) ROC areas based on 3-h accumulated precipitation aggregated over all 35 forecasts as a function of forecast hour for verification over the central and eastern CONUS regions, spanning 105°–87°W and the CONUS east of 87°W, respectively (Fig. 1). Shaded gray cells for forecast hours 15–75 indicate instances when the 3-km MPAS ensemble had statistically significantly better scores than the 15-km MPAS ensemble at the 95% level. White cells mean the 3-km MPAS ensemble was not statistically significantly better than the 15-km MPAS ensemble at the 95% level. Statistically significant differences after 75 h were rare, so for hours 78–132, those forecast periods with statistically significantly better 3-km forecasts are explicitly listed. Forecast hours represent ending times of 3-h accumulation periods (e.g., a value of 24 is for 3-h accumulated precipitation between 21 and 24 h).

3-km convection for precipitation forecasting could be realized.

7. Case study

Forecasts of 3-h accumulated precipitation between 2100 UTC 16 May and 0000 UTC 17 May 2017 illustrate several properties of the bulk performance statistics and provide an opportunity to subjectively evaluate other forecast properties not explicitly measured by probabilistic objective metrics, like convective orientation and mode. During this 3-h period, severe weather and heavy precipitation was ongoing over portions of the Great Plains (Fig. 15a) ahead of a potent upper-level trough and surface cold front and dryline, with supercells in portions of western Oklahoma, the eastern Texas Panhandle, and southern Kansas, while a complex comprising mostly linear features developed farther north.

Individual GEFS members' 120-h forecasts of 3-h accumulated precipitation ending 0000 UTC 17 May failed to reproduce the $25.0 \text{ mm (3 h)}^{-1}$ rates that were observed (Figs. 15j–m), reflecting areal coverage statistics (Figs. 4e,f). Conversely, all 15- and 3-km MPAS ensemble members suggested precipitation $\geq 20 \text{ mm (3 h)}^{-1}$ somewhere over the central plains while conveying uncertainty regarding placement (Figs. 15b–i). For example, some members incorrectly focused precipitation in central and eastern Oklahoma (Figs. 15c,e,g,i) while others suggested rainfall farther west (Figs. 15b,d,h), closer to where precipitation actually occurred (Fig. 15a). There were also differences regarding precipitation location and areal extent in Nebraska and northern Texas.

At 120 h, precipitation locations in corresponding 3- and 15-km MPAS ensemble members were broadly similar, with differences primarily concerning convective orientation and mode. Generally, whereas 15-km

FIG. 15. (a) Observed and (b)–(m) 120-h forecasts of 3-h accumulated precipitation (mm) valid at 0000 UTC 17 May 2017 for members 1, 3, 6, and 10 of the (b)–(e) 3-km MPAS ensemble (MPAS-3km), (f)–(i) 15-km MPAS ensemble (MPAS-15km), and (j)–(m) GEFS. Forecasts in (b)–(m) were initialized at 0000 UTC 12 May 2017. Solid black contours enclose grid points meeting or exceeding the 99th percentile threshold as computed over the entire verification region, which spans a larger area than pictured. Nearly all points with precipitation $\geq 1.0 \text{ mm (3 h)}^{-1}$ exceeded the 95th percentile threshold.

MPAS members produced broad features, 3-km MPAS members provided more detail, including discrete storms over parts of Oklahoma and the eastern Texas panhandle (Figs. 15b–e) similar to those observed, despite sometimes large position errors. Additionally,

diverse modes were produced over Kansas, with some 3-km members predicting isolated structures (Figs. 15c,d) while others suggested linear features or complexes (Figs. 15b,e). Accompanying the more detailed structures, 3-km features regularly had SW–NE orientations,

similar to observations, while features in many 15-km MPAS members had S–N orientations.

The 72-h GEFS forecast was improved compared to its 120-h forecast, with several members indicating areas of heavier precipitation (Figs. 16j–m). Both MPAS ensembles had less spread compared to their 120-h forecasts (reflecting Fig. 12) and generally smaller false alarm errors in eastern Oklahoma and northern Texas (Figs. 16b–i). Again, 3-km members had more detailed structures, especially in Oklahoma, which were similar to those observed.

Forecasts continued to improve as lead time decreased to 24 h: GEFS positions of heavy rainfall improved (Figs. 17j–m) and 3-km features were more accurately placed (Figs. 17b–e). As at longer lead times, 15-km forecasts had broader structures that did not explicitly suggest isolated storms in Texas and Oklahoma (Figs. 17f–i) whereas the 3-km forecasts did. At 24 h, the 3-km forecasts visually seem to agree best with observations, although some errors remained. It is interesting that member 10 in all three ensembles produced the heaviest precipitation over Kansas, which likely reflects influence of common initial conditions at this relatively early forecast time (Figs. 17e,i,m).

NEPs computed with a 100-km neighborhood length scale for the 95th percentile threshold summarize aspects regarding spatial placement (Fig. 18). At 120 h, both MPAS ensembles had substantial false alarm areas in northern Texas and central–eastern Oklahoma, while GEFS false alarm areas were notably smaller in these regions (Figs. 18a–c). However, at this time, over central Kansas, where observed events occurred, GEFS NEPs were <20% while the two MPAS ensembles had NEPs >40%. Thus, subjectively determining the best probabilistic forecast for the 95th percentile 120-h forecast is difficult; while MPAS NEPs were higher in and around areas where observed events occurred, the MPAS ensembles had larger false alarm areas than the GEFS, and the two MPAS forecasts were fairly similar. These findings reflect aggregate objective statistics showing differences between the three ensembles were usually statistically insignificant around 120 h, especially over the central CONUS (e.g., Fig. 14).

As lead time decreased, MPAS forecasts improved: false alarm areas shrank and probabilities generally increased over central Kansas (Figs. 18d–i). By 24 h, 3-km MPAS NEPs were highest over Kansas, indicating the best correspondence with observations, while the GEFS and 15-km MPAS NEPs were lower (Figs. 18g–i). All ensembles struggled with isolated convection over the eastern Texas panhandle and extreme western Oklahoma, although the 3-km MPAS

ensemble had the highest NEPs in the vicinity of those observed events.

In summary, for this event, 5-day forecasts from 3-km MPAS ensemble members suggested potential for heavy rainfall, despite uncertain details. As the event drew nearer, probabilities generally increased in regions where heavy rainfall occurred and decreased in regions not receiving precipitation, with 3-km advantages becoming most apparent by 24 h, reflecting objective statistics (section 5). Furthermore, 3-km members provided explicit information about convective mode that sometimes resembled observed structures, while the coarser-resolution ensembles were unable to provide storm-scale details.

8. Summary and future directions

Thirty-five 132-h, 10-member ensemble forecasts from the MPAS model on both quasi-uniform 15-km and variable-resolution 15-/3-km global meshes (Fig. 1) were verified over the CONUS with a focus on precipitation. Forecasts from NCEP's operational GEFS were also evaluated. Collective results suggested both benefits and potential limitations of configuring MPAS with a 3-km mesh refinement region for 5-day forecasts:

- 1) Both the diurnal cycle and climatology of precipitation were best represented by the 3-km MPAS ensemble throughout the 132-h forecasts (Figs. 4–6).
- 2) Using percentile-based objective verification, which removed bias and permitted robust assessments of spatial placement, ROC areas, FSSs, and attributes statistics indicated the 3-km MPAS ensemble had better probabilistic precipitation forecast skill and reliability than the GEFS and 15-km MPAS ensemble over the first 48 h (Figs. 7, 8, 10, 11). The 3-km ensemble had excellent day 1 reliability.
- 3) After 48 h, percentile-based objective verification revealed that probabilistic precipitation forecasts from all three ensembles were overconfident, and differences between the 15- and 3-km MPAS ensembles were usually statistically insignificant. The similar verification scores for the 15- and 3-km MPAS ensembles after 48 h may have been due to greater accumulation and upscale growth of small-scale errors in the 3-km forecasts that limited 3-km forecast quality (e.g., Fig. 13). Both MPAS ensembles generally outperformed the GEFS throughout the 132-h forecasts.
- 4) Variance of 3-h accumulated precipitation grew fastest in the 3-km MPAS ensemble over the first ~30 h, reflecting rapid upscale growth of storm-scale errors (Fig. 12). However, thereafter, the 3- and 15-km MPAS ensembles had similar variance growth rates.

FIG. 16. As in Fig. 15, but for the 72-h forecast initialized 0000 UTC 14 May 2017, also valid at 0000 UTC 17 May 2017.

5) Individual 3-km ensemble members explicitly predicted convective mode throughout the 132-h forecasts whereas coarser-resolution members did not (Figs. 15–17). This finding was expected and due to inherent benefits of convection-allowing resolution (e.g., Kain et al. 2006; Weisman et al. 2008; Schwartz et al. 2009). Nonetheless, information and uncertainty about convective mode in the 3–5-day forecast range is potentially valuable.

Overall, these results provide somewhat ambiguous conclusions regarding necessity for 3–5-day ensemble forecasts with a variable-resolution global model featuring a large convection-allowing mesh-refinement region. There is no evidence convection-allowing resolution improved representation of 3–5-day synoptic-scale patterns that modulate where precipitation occurs, and objective statistics revealed few 3-km advantages after 48 h. Thus, forecasters relying on environmental

FIG. 17. As in Fig. 15, but for the 24-h forecast initialized 0000 UTC 16 May 2017, also valid at 0000 UTC 17 May 2017.

parameters to provide medium-range forecast information about rainfall amounts and convective mode (e.g., Thompson et al. 2012) might be comfortable consulting convection-parameterizing ensembles and find the extra expense (Table 2) to attain convection-allowing resolution unwarranted. Conversely, if explicit 3–5-day numerical guidance about convective mode and precipitation intensity is desired, there is no

substitute for refining grid spacing to convection-allowing resolution.

Because a small sample (35 cases) was examined, the ensembles were small (10 members), and absence of physically based stochastic perturbations restricted MPAS ensemble spread, these results have limitations and should be interpreted cautiously as an early examination of global CAE performance. Furthermore, while

FIG. 18. Neighborhood ensemble probabilities (NEPs; %) of 3-h accumulated precipitation exceeding 95th percentile thresholds computed with a 100-km neighborhood length scale valid at 0000 UTC 17 May 2017 for (a)–(c) 120-, (d)–(f) 72-, and (g)–(i) 24-h forecasts initialized 0000 UTC 12, 14, and 16 May 2017, respectively, for the 10-member (a),(d),(g) 3-km MPAS ensemble (MPAS-3km), (b),(e),(h) 15-km MPAS ensemble (MPAS-15km), and (c),(f),(i) GEFS. Points where observed (i.e., ST4) precipitation exceeded its 95th percentile are overlaid and enclosed by black contours.

evaluating microphysical, thermodynamic, and dynamic aspects of the forecasts is outside the scope of this paper, these additional investigations would be useful to better understand model performance and convergence of 15- and 3-km MPAS ensemble forecast skill after ~ 48 h.

Indeed, much more research is required to understand global CAE behavior, strengths, weaknesses, and potential operational use. For example, further studies should rigorously verify 3–5-day CAE forecasts of convective mode, structure, orientation, and morphology to determine whether storm-scale information at these

forecast ranges represents valuable guidance even if displacement errors are large. Moreover, when computing resources permit, spread and skill in *global, quasi-uniform 3-km CAEs* should be assessed to determine whether CAEs employing variable-resolution meshes to achieve convection-allowing resolution have major limitations. It would also be beneficial to investigate how forecasts downstream of mesh-refinement regions are impacted by the enhanced upstream resolution. Furthermore, using spectral analysis techniques to quantify error growth characteristics and

TABLE 2. Computational cost to produce 132-h, 10-member MPAS ensemble forecasts. The total cost was the product of number of processor cores (N), wall clock time (W), and ensemble size (E).

Ensemble	No. of cores required for one ensemble member (N)	Approximate wall clock time (W) (h)	Ensemble size (E)	Total cost ($N \times W \times E$)
Quasi-uniform 15-km MPAS ensemble	2340	2.25	10	52 650
Variable-resolution 15-/3-km MPAS ensemble	9720	4.25	10	413 100

scale interactions in variable-resolution global models with convection-allowing areas would be valuable.

Finally, future work should examine whether running global CAEs instead of limited-area CAEs is justified, especially if the global model uses a variable-resolution mesh. Although MPAS was designed as a global NWP model, a limited-area version is under development (Skamarock et al. 2018) that will permit opportunities to cleanly assess whether global, variable-resolution, medium-range forecasts improve upon less-expensive limited-area forecasts for convection-allowing applications.

Acknowledgments. Thanks to David Ahijevych (NCAR/MMM) for assistance plotting MPAS output; Laura Fowler, Morris Weisman, Falko Judt, and Bill Skamarock (NCAR/MMM) for internal reviews; and the entire MPAS development team for their efforts. Three anonymous reviewers provided useful comments that improved this paper. This work was partially funded by NCAR's Short-term Explicit Prediction (STEP) program. Computations were performed on NCAR's Cheyenne supercomputer (Computational and Information Systems Laboratory 2017). NCAR is sponsored by the National Science Foundation.

REFERENCES

- Accadia, C., S. Mariani, M. Casaioli, A. Lavagnini, and A. Speranza, 2003: Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbor average method on high-resolution verification grids. *Wea. Forecasting*, **18**, 918–932, [https://doi.org/10.1175/1520-0434\(2003\)018<0918:SOPFSS>2.0.CO;2](https://doi.org/10.1175/1520-0434(2003)018<0918:SOPFSS>2.0.CO;2).
- Baldwin, M. E., and J. S. Kain, 2006: Sensitivity of several performance measures to displacement error, bias, and event frequency. *Wea. Forecasting*, **21**, 636–648, <https://doi.org/10.1175/WAF933.1>.
- Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2).
- Chen, F., and J. Dudhia, 2001: Coupling an advanced land-surface–hydrology model with the Penn State–NCAR MM5 modeling system. Part I: Model description and implementation. *Mon. Wea. Rev.*, **129**, 569–585, [https://doi.org/10.1175/1520-0493\(2001\)129<0569:CAALSH>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<0569:CAALSH>2.0.CO;2).
- Clark, A. J., W. A. Gallus Jr., and T.-C. Chen, 2007: Comparison of the diurnal precipitation cycle in convection-resolving and non-convection-resolving mesoscale models. *Mon. Wea. Rev.*, **135**, 3456–3473, <https://doi.org/10.1175/MWR3467.1>.
- , —, 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, <https://doi.org/10.1175/2009WAF2222222.1>.
- , —, and M. L. Weisman, 2010a: Neighborhood-based verification of precipitation forecasts from convection-allowing NCAR WRF Model simulations and the operational NAM. *Wea. Forecasting*, **25**, 1495–1509, <https://doi.org/10.1175/2010WAF2222404.1>.
- , —, M. Xue, and F. Kong, 2010b: Growth of spread in convection-allowing and convection-parameterizing ensembles. *Wea. Forecasting*, **25**, 594–612, <https://doi.org/10.1175/2009WAF2222318.1>.
- , 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, <https://doi.org/10.1175/2010MWR3624.1>.
- , and Coauthors, 2012: An overview of the 2010 Hazardous Weather Testbed experimental forecasting program spring experiment. *Bull. Amer. Meteor. Soc.*, **93**, 55–74, <https://doi.org/10.1175/BAMS-D-11-00040.1>.
- , and Coauthors, 2018: The Community Leveraged Unified Ensemble (CLUE) in the 2016 NOAA/Hazardous Weather Testbed spring forecasting experiment. *Bull. Amer. Meteor. Soc.*, **99**, 1433–1448, <https://doi.org/10.1175/BAMS-D-16-0309.1>.
- Clark, P., N. Roberts, H. Lean, S. P. Ballard, and C. Charlton-Perez, 2016: Convection-permitting models: A step-change in rainfall forecasting. *Meteor. Appl.*, **23**, 165–181, <https://doi.org/10.1002/met.1538>.
- Computational and Information Systems Laboratory, 2017: Cheyenne: HPE/SGI ICE XA system (NCAR community computing). National Center for Atmospheric Research, accessed 17 July 2019, <https://doi.org/10.5065/D6RX99HX>.
- Davis, C. A., D. A. Ahijevych, W. Wang, and W. C. Skamarock, 2016: Evaluating medium-range tropical cyclone forecasts in uniform- and variable-resolution global models. *Mon. Wea. Rev.*, **144**, 4141–4160, <https://doi.org/10.1175/MWR-D-16-0021.1>.
- Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, <https://doi.org/10.1002/qj.828>.
- Dey, S. R., G. Leoncini, N. M. Roberts, R. S. Plant, and S. Migliorini, 2014: A spatial view of ensemble spread in convection permitting ensembles. *Mon. Wea. Rev.*, **142**, 4091–4107, <https://doi.org/10.1175/MWR-D-14-00172.1>.
- Done, J., C. A. Davis, and M. L. Weisman, 2004: The next generation of NWP: Explicit forecasts of convection using the Weather Research and Forecasting (WRF) Model. *Atmos. Sci. Lett.*, **5**, 110–117, <https://doi.org/10.1002/asl.72>.
- Duc, L., K. Saito, and H. Seko, 2013: Spatial-temporal fractions verification for high-resolution ensemble forecasts. *Tellus*, **65A**, 18171, <https://doi.org/10.3402/tellusa.v65i0.18171>.

- Durran, D. R., and M. Gingrich, 2014: Atmospheric predictability: Why butterflies are not important. *J. Atmos. Sci.*, **71**, 2476–2488, <https://doi.org/10.1175/JAS-D-14-0007.1>.
- , and J. A. Weyn, 2016: Thunderstorms do not get butterflies. *Bull. Amer. Meteor. Soc.*, **97**, 237–243, <https://doi.org/10.1175/BAMS-D-15-00070.1>.
- Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461–2480, [https://doi.org/10.1175/1520-0493\(2001\)129<2461:AOAPMS>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<2461:AOAPMS>2.0.CO;2).
- , 2008: Fuzzy verification of high resolution gridded forecasts: A review and proposed framework. *Meteor. Appl.*, **15**, 51–64, <https://doi.org/10.1002/met.25>.
- , 2009: Neighborhood verification: A strategy for rewarding close forecasts. *Wea. Forecasting*, **24**, 1498–1510, <https://doi.org/10.1175/2009WAF2222251.1>.
- Evans, C., D. F. Van Dyke, and T. Lericos, 2014: How do forecasters utilize output from a convection-permitting ensemble forecast system? Case study of a high-impact precipitation event. *Wea. Forecasting*, **29**, 466–486, <https://doi.org/10.1175/WAF-D-13-00064.1>.
- Fowler, L. D., W. C. Skamarock, G. A. Grell, S. R. Freitas, and M. G. Duda, 2016: Analyzing the Grell–Freitas convection scheme from hydrostatic to nonhydrostatic scales within a global model. *Mon. Wea. Rev.*, **144**, 2285–2306, <https://doi.org/10.1175/MWR-D-15-0311.1>.
- Fox-Rabinovitz, M., J. Côté, B. Dugas, M. Déqué, and J. McGregor, 2006: Variable resolution general circulation models: Stretched-grid model intercomparison project (SGMIP). *J. Geophys. Res.*, **111**, D16104, <https://doi.org/10.1029/2005JD006520>.
- Gallo, B. T., and Coauthors, 2017: Breaking new ground in severe weather prediction: The 2015 NOAA/Hazardous Weather Testbed spring forecasting experiment. *Wea. Forecasting*, **32**, 1541–1568, <https://doi.org/10.1175/WAF-D-16-0178.1>.
- Gallus, W. A., Jr., 2002: Impact of verification grid-box size on warm-season QPF skill measures. *Wea. Forecasting*, **17**, 1296–1302, [https://doi.org/10.1175/1520-0434\(2002\)017<1296:IOVGBS>2.0.CO;2](https://doi.org/10.1175/1520-0434(2002)017<1296:IOVGBS>2.0.CO;2).
- Gebhardt, C., S. E. Theis, M. Paulat, and Z. Ben Bouallègue, 2011: Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variation of lateral boundaries. *Atmos. Res.*, **100**, 168–177, <https://doi.org/10.1016/j.atmosres.2010.12.008>.
- Geerts, B., and Coauthors, 2017: The 2015 Plains Elevated Convection at Night (PECAN) field project. *Bull. Amer. Meteor. Soc.*, **98**, 767–786, <https://doi.org/10.1175/BAMS-D-15-00257.1>.
- Gilleland, E., 2010: Confidence intervals for forecast verification. NCAR Tech. Note NCAR/TN-479+STR, 71 pp., <https://doi.org/10.5065/D6WD3XJM>.
- Gowan, T. M., W. J. Steenburgh, and C. S. Schwartz, 2018: Validation of mountain precipitation forecasts from the convection-permitting NCAR ensemble and operational forecast systems over the western United States. *Wea. Forecasting*, **33**, 739–765, <https://doi.org/10.1175/WAF-D-17-0144.1>.
- Grell, G. A., and S. R. Freitas, 2014: A scale and aerosol aware stochastic convective parameterization for weather and air quality modeling. *Atmos. Chem. Phys.*, **14**, 5233–5250, <https://doi.org/10.5194/acp-14-5233-2014>.
- Ha, S., C. Snyder, W. C. Skamarock, J. L. Anderson, and N. Collins, 2017: Ensemble Kalman filter data assimilation for the Model for Prediction Across Scales (MPAS). *Mon. Wea. Rev.*, **145**, 4673–4692, <https://doi.org/10.1175/MWR-D-17-0145.1>.
- Hagelin, S., J. Son, R. Swinbank, A. McCabe, N. Roberts, and W. Tennant, 2017: The Met Office convective-scale ensemble, MOGREPS-UK. *Quart. J. Roy. Meteor. Soc.*, **143**, 2846–2861, <https://doi.org/10.1002/qj.3135>.
- Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167, [https://doi.org/10.1175/1520-0434\(1999\)014<0155:HTFENP>2.0.CO;2](https://doi.org/10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2).
- Harris, L. M., and S.-J. Lin, 2013: A two-way nested global-regional dynamical core on the cubed-sphere grid. *Mon. Wea. Rev.*, **141**, 283–306, <https://doi.org/10.1175/MWR-D-11-00201.1>.
- , —, and C. Tu, 2016: High-resolution climate simulations using GFDL HiRAM with a stretched global grid. *J. Climate*, **29**, 4293–4314, <https://doi.org/10.1175/JCLI-D-15-0389.1>.
- Heinzeller, D., M. G. Duda, and H. Kunstmann, 2016: Towards convection-resolving, global atmospheric simulations with the Model for Prediction Across Scales (MPAS) v3.1: An extreme scaling experiment. *Geosci. Model Dev.*, **9**, 77–110, <https://doi.org/10.5194/gmd-9-77-2016>.
- Hohenegger, C., and C. Schär, 2007: Predictability and error growth dynamics in cloud-resolving models. *J. Atmos. Sci.*, **64**, 4467–4478, <https://doi.org/10.1175/2007JAS2143.1>.
- Hou, D., Z. Toth, and Y. Zhu, 2006: A stochastic parameterization scheme within NCEP global ensemble forecast system. *18th Conf. on Probability and Statistics in the Atmospheric Sciences*, Atlanta, GA, Amer. Meteor. Soc., 4.5, <https://ams.confex.com/ams/pdfpapers/101401.pdf>.
- Iacono, M. J., J. S. Delamere, E. J. Mlawer, M. W. Shephard, S. A. Clough, and W. D. Collins, 2008: Radiative forcing by long-lived greenhouse gases: Calculations with the AER radiative transfer models. *J. Geophys. Res.*, **113**, D13103, <https://doi.org/10.1029/2008JD009944>.
- Iyer, E. R., A. J. Clark, M. Xue, and F. Kong, 2016: A comparison of 36–60-h precipitation forecasts from convection-allowing and convection-parameterizing ensembles. *Wea. Forecasting*, **31**, 647–661, <https://doi.org/10.1175/WAF-D-15-0143.1>.
- Jones, P. W., 1999: First- and second-order conservative remapping schemes for grids in spherical coordinates. *Mon. Wea. Rev.*, **127**, 2204–2210, [https://doi.org/10.1175/1520-0493\(1999\)127<2204:FASOCR>2.0.CO;2](https://doi.org/10.1175/1520-0493(1999)127<2204:FASOCR>2.0.CO;2).
- Judt, F., 2018: Insights into atmospheric predictability through global convection-permitting model simulations. *J. Atmos. Sci.*, **75**, 1477–1497, <https://doi.org/10.1175/JAS-D-17-0343.1>.
- Kain, J. S., S. J. Weiss, J. J. Levit, M. E. Baldwin, and D. R. Bright, 2006: Examination of convection-allowing configurations of the WRF Model for the prediction of severe convective weather: The SPC/NSSL Spring Program 2004. *Wea. Forecasting*, **21**, 167–181, <https://doi.org/10.1175/WAF906.1>.
- Klasa, C., M. Arpagaus, A. Walser, and H. Wernli, 2018: An evaluation of the convection-permitting ensemble COSMO-E for three contrasting precipitation events in Switzerland. *Quart. J. Roy. Meteor. Soc.*, **144**, 744–764, <https://doi.org/10.1002/qj.3245>.
- Kleist, D. T., and K. Ide, 2015a: An OSSE-based evaluation of hybrid variational-ensemble data assimilation for the NCEP GFS. Part I: System description and 3D-hybrid results. *Mon. Wea. Rev.*, **143**, 433–451, <https://doi.org/10.1175/MWR-D-13-00351.1>.
- , and —, 2015b: An OSSE-based evaluation of hybrid variational-ensemble data assimilation for the NCEP GFS. Part II: 4D-EnVar and hybrid variants. *Mon. Wea. Rev.*, **143**, 452–470, <https://doi.org/10.1175/MWR-D-13-00350.1>.
- Klemp, J. B., 2011: A terrain-following coordinate with smoothed coordinate surfaces. *Mon. Wea. Rev.*, **139**, 2163–2169, <https://doi.org/10.1175/MWR-D-10-05046.1>.

- Lin, Y., and K. E. Mitchell, 2005: The NCEP stage II/IV hourly precipitation analyses: Development and applications. *19th Conf. on Hydrology*, San Diego, CA, Amer. Meteor. Soc., 1,2, <https://ams.confex.com/ams/pdfpapers/83847.pdf>.
- Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.*, **20**, 130–141, [https://doi.org/10.1175/1520-0469\(1963\)020<0130:DNF>2.0.CO;2](https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2).
- , 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, **21**, 289–307, <https://doi.org/10.3402/tellusa.v21i3.10086>.
- Mason, I. B., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.*, **30**, 291–303.
- Mason, S. J., and N. E. Graham, 2002: Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quart. J. Roy. Meteor. Soc.*, **128**, 2145–2166, <https://doi.org/10.1256/003590002320603584>.
- Mittermaier, M., 2018: How interpolation and resolution can affect verification scores: A study based on the Fractions Skill Score. *Meteor. Z.*, <https://doi.org/10.1127/metz/2018/0890>, in press.
- , and N. Roberts, 2010: Intercomparison of spatial forecast methods: Identifying skillful spatial scales using the fractions skill score. *Wea. Forecasting*, **25**, 343–354, <https://doi.org/10.1175/2009WAF2222260.1>.
- , —, and S. A. Thompson, 2013: A long-term assessment of precipitation forecast skill using the fractions skill score. *Meteor. Appl.*, **20**, 176–186, <https://doi.org/10.1002/met.296>.
- Miura, H., M. Satoh, T. Nasuno, A. T. Noda, and K. Oouchi, 2007: A Madden-Julian oscillation event realistically simulated by a global cloud-resolving model. *Science*, **318**, 1763–1765, <https://doi.org/10.1126/science.1148443>.
- Miyamoto, Y., Y. Kajikawa, R. Yoshida, T. Yamaura, H. Yashiro, and H. Tomita, 2013: Deep moist atmospheric convection in a subkilometer global simulation. *Geophys. Res. Lett.*, **40**, 4922–4926, <https://doi.org/10.1002/grl.50944>.
- , T. Yamaura, R. Yoshida, H. Yashiro, H. Tomita, and Y. Kajikawa, 2016: Precursors of deep moist convection in a subkilometer global simulation. *J. Geophys. Res. Atmos.*, **121**, 12 080–12 088, <https://doi.org/10.1002/2016JD024965>.
- Mlawer, E. J., S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough, 1997: Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. *J. Geophys. Res.*, **102**, 16 663–16 682, <https://doi.org/10.1029/97JD00237>.
- Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281–293, [https://doi.org/10.1175/1520-0434\(1993\)008<0281:WIAGFA>2.0.CO;2](https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2).
- Nakanishi, M., and H. Niino, 2009: Development of an improved turbulence closure model for the atmospheric boundary layer. *J. Meteor. Soc. Japan*, **87**, 895–912, <https://doi.org/10.2151/jmsj.87.895>.
- Nelson, B. R., O. P. Prat, D.-J. Seo, and E. Habib, 2016: Assessment and implications of NCEP Stage IV quantitative precipitation estimates for product intercomparisons. *Wea. Forecasting*, **31**, 371–394, <https://doi.org/10.1175/WAF-D-14-00112.1>.
- Nutter, P., D. Stensrud, and M. Xue, 2004: Effects of coarsely resolved and temporally interpolated lateral boundary conditions on the dispersion of limited-area ensemble forecasts. *Mon. Wea. Rev.*, **132**, 2358–2377, [https://doi.org/10.1175/1520-0493\(2004\)132<2358:EOCRAT>2.0.CO;2](https://doi.org/10.1175/1520-0493(2004)132<2358:EOCRAT>2.0.CO;2).
- Peralta, C., Z. B. Bouallègue, S. E. Theis, C. Gebhardt, and M. Buchhold, 2012: Accounting for initial condition uncertainties in COSMO-DE-EPS. *J. Geophys. Res.*, **117**, D07108, <https://doi.org/10.1029/2011JD016581>.
- Potvin, C. K., E. M. Murillo, M. L. Flora, and D. M. Wheatley, 2017: Sensitivity of supercell simulations to initial-condition resolution. *J. Atmos. Sci.*, **74**, 5–26, <https://doi.org/10.1175/JAS-D-16-0098.1>.
- Powers, J. G., and Coauthors, 2017: The Weather Research and Forecasting Model: Overview, system efforts, and future directions. *Bull. Amer. Meteor. Soc.*, **98**, 1717–1737, <https://doi.org/10.1175/BAMS-D-15-00308.1>.
- Putman, W. M., and M. Suarez, 2011: Cloud-system resolving simulations with the NASA Goddard Earth Observing System global atmospheric model (GEOS-5). *Geophys. Res. Lett.*, **38**, L16809, <https://doi.org/10.1029/2011GL048438>.
- Raynaud, L., and F. Bouttier, 2017: The impact of horizontal resolution and ensemble size for convective-scale probabilistic forecasts. *Quart. J. Roy. Meteor. Soc.*, **143**, 3037–3047, <https://doi.org/10.1002/qj.3159>.
- 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, <https://doi.org/10.1175/2007MWR2123.1>.
- Satoh, M., and Coauthors, 2017: Outcomes and challenges of global high-resolution non-hydrostatic atmospheric simulations using the K computer. *Prog. Earth Planet. Sci.*, **4**, 13, <https://doi.org/10.1186/s40645-017-0127-8>.
- , T. Matsuno, H. Tomita, H. Miura, T. Nasuno, and S. Iga, 2008: Nonhydrostatic icosahedral atmospheric model (NICAM) for global cloud resolving simulations. *J. Comput. Phys.*, **227**, 3486–3514, <https://doi.org/10.1016/j.jcp.2007.02.006>.
- Schellander-Gorgas, T., Y. Wang, F. Meier, F. Weidle, C. Wittmann, and A. Kann, 2017: On the forecast skills of a convection-permitting ensemble. *Geosci. Model Dev.*, **10**, 35–56, <https://doi.org/10.5194/gmd-10-35-2017>.
- Schwartz, C. S., 2014: Reproducing the September 2013 record-breaking rainfall over the Colorado Front Range with high-resolution WRF forecasts. *Wea. Forecasting*, **29**, 393–402, <https://doi.org/10.1175/WAF-D-13-00136.1>.
- , and R. A. Sobash, 2017: Generating probabilistic forecasts from convection-allowing ensembles using neighborhood approaches: A review and recommendations. *Mon. Wea. Rev.*, **145**, 3397–3418, <https://doi.org/10.1175/MWR-D-16-0400.1>.
- , and Coauthors, 2009: Next-day convection-allowing WRF Model guidance: A second look at 2-km versus 4-km grid spacing. *Mon. Wea. Rev.*, **137**, 3351–3372, <https://doi.org/10.1175/2009MWR2924.1>.
- , and Coauthors, 2010: Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Wea. Forecasting*, **25**, 263–280, <https://doi.org/10.1175/2009WAF2222267.1>.
- , G. S. Romine, K. R. Smith, and M. L. Weisman, 2014: Characterizing and optimizing precipitation forecasts from a convection-permitting ensemble initialized by a mesoscale ensemble Kalman filter. *Wea. Forecasting*, **29**, 1295–1318, <https://doi.org/10.1175/WAF-D-13-00145.1>.
- , —, M. L. Weisman, R. A. Sobash, K. R. Fossell, K. W. Manning, and S. B. Trier, 2015a: A real-time convection-allowing ensemble prediction system initialized by mesoscale ensemble Kalman filter analyses. *Wea. Forecasting*, **30**, 1158–1181, <https://doi.org/10.1175/WAF-D-15-0013.1>.
- , —, R. A. Sobash, K. R. Fossell, and M. L. Weisman, 2015b: NCAR's experimental real-time convection-allowing

- ensemble prediction system. *Wea. Forecasting*, **30**, 1645–1654, <https://doi.org/10.1175/WAF-D-15-0103.1>.
- , —, —, —, and —, 2019: NCAR's real-time convection-allowing ensemble project. *Bull. Amer. Meteor. Soc.*, **100**, 321–343, <https://doi.org/10.1175/BAMS-D-17-0297.1>.
- Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-475+STR, 113 pp., <https://doi.org/10.5065/D68S4MVH>.
- , J. B. Klemp, M. G. Duda, L. D. Fowler, S.-H. Park, and T. D. Ringler, 2012: A multiscale nonhydrostatic atmospheric model using centroidal Voronoi tessellations and C-grid staggering. *Mon. Wea. Rev.*, **140**, 3090–3105, <https://doi.org/10.1175/MWR-D-11-00215.1>.
- , S.-H. Park, J. B. Klemp, and C. Snyder, 2014: Atmospheric kinetic energy spectra from global high-resolution nonhydrostatic simulations. *J. Atmos. Sci.*, **71**, 4369–4381, <https://doi.org/10.1175/JAS-D-14-0114.1>.
- , M. G. Duda, S. Ha, and S. Park, 2018: Limited-area atmospheric modeling using an unstructured mesh. *Mon. Wea. Rev.*, **146**, 3445–3460, <https://doi.org/10.1175/MWR-D-18-0155.1>.
- Sobash, R. A., C. S. Schwartz, G. S. Romine, K. R. Fossell, and M. L. Weisman, 2016a: Severe weather prediction using storm surrogates from an ensemble forecasting system. *Wea. Forecasting*, **31**, 255–271, <https://doi.org/10.1175/WAF-D-15-0138.1>.
- , G. S. Romine, C. S. Schwartz, D. J. Gagne II, and M. L. Weisman, 2016b: Explicit forecasts of low-level rotation from convection-allowing models for next-day tornado prediction. *Wea. Forecasting*, **31**, 1591–1614, <https://doi.org/10.1175/WAF-D-16-0073.1>.
- Stensrud, D. J., and Coauthors, 2009: Convective-scale warn-on-forecast system: A vision for 2020. *Bull. Amer. Meteor. Soc.*, **90**, 1487–1499, <https://doi.org/10.1175/2009BAMS2795.1>.
- Surcel, M., I. Zawadzki, M. K. Yau, M. Xue, and F. Kong, 2017: More on the scale dependence of the predictability of precipitation patterns: Extension to the 2009–13 CAPS Spring Experiment ensemble forecasts. *Mon. Wea. Rev.*, **145**, 3625–3646, <https://doi.org/10.1175/MWR-D-16-0362.1>.
- Theis, S. E., A. Hense, and U. Damrath, 2005: Probabilistic precipitation forecasts from a deterministic model: A pragmatic approach. *Meteor. Appl.*, **12**, 257–268, <https://doi.org/10.1017/S1350482705001763>.
- Thompson, G., P. R. Field, R. M. Rasmussen, and W. D. Hall, 2008: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization. *Mon. Wea. Rev.*, **136**, 5095–5115, <https://doi.org/10.1175/2008MWR2387.1>.
- Thompson, R. L., B. T. Smith, J. S. Grams, A. R. Dean, and C. Broyles, 2012: Convective modes for significant severe thunderstorms in the contiguous United States. Part II: Supercell and QLCS tornado environments. *Wea. Forecasting*, **27**, 1136–1154, <https://doi.org/10.1175/WAF-D-11-00116.1>.
- Tomita, H., 2008: A stretched grid on a sphere by new grid transformation. *J. Meteor. Soc. Japan*, **86A**, 107–119, <https://doi.org/10.2151/jmsj.86A.107>.
- Wang, X., and T. Lei, 2014: GSI-based four-dimensional ensemble-variational (4DEnsVar) data assimilation: Formulation and single-resolution experiments with real data for the NCEP Global Forecast System. *Mon. Wea. Rev.*, **142**, 3303–3325, <https://doi.org/10.1175/MWR-D-13-00303.1>.
- Warner, T. T., R. A. Peterson, and R. E. Treadon, 1997: A tutorial on lateral boundary conditions as a basic and potentially serious limitation to regional numerical weather prediction. *Bull. Amer. Meteor. Soc.*, **78**, 2599–2617, [https://doi.org/10.1175/1520-0477\(1997\)078<2599:ATOLBC>2.0.CO;2](https://doi.org/10.1175/1520-0477(1997)078<2599:ATOLBC>2.0.CO;2).
- Weisman, M. L., C. Davis, W. Wang, K. W. Manning, and J. B. Klemp, 2008: Experiences with 0–36-h explicit convective forecasts with the WRF-ARW Model. *Wea. Forecasting*, **23**, 407–437, <https://doi.org/10.1175/2007WAF2007005.1>.
- Weyn, J. A., and D. R. Durran, 2017: The dependence of the predictability of mesoscale convective systems on the horizontal scale and amplitude of initial errors in idealized simulations. *J. Atmos. Sci.*, **74**, 2191–2210, <https://doi.org/10.1175/JAS-D-17-0006.1>.
- , and —, 2019: The scale dependence of initial-condition sensitivities in simulations of convective systems over the southeastern United States. *Quart. J. Roy. Meteor. Soc.*, <https://doi.org/10.1002/qj.3367>, in press.
- Wheatley, D. M., K. H. Knopfmeier, T. A. Jones, and G. J. Creager, 2015: Storm-scale data assimilation and ensemble forecasting with the NSSL experimental Warn-on-Forecast system. Part I: Radar data experiments. *Wea. Forecasting*, **30**, 1795–1817, <https://doi.org/10.1175/WAF-D-15-0043.1>.
- Whitaker, J. S., and T. M. Hamill, 2002: Ensemble data assimilation without perturbed observations. *Mon. Wea. Rev.*, **130**, 1913–1924, [https://doi.org/10.1175/1520-0493\(2002\)130<1913:EDAWPO>2.0.CO;2](https://doi.org/10.1175/1520-0493(2002)130<1913:EDAWPO>2.0.CO;2).
- Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. Elsevier, 676 pp.
- Wolff, J. K., M. Harrold, T. Fowler, J. H. Gotway, L. Nance, and B. G. Brown, 2014: Beyond the basics: Evaluating model-based precipitation forecasts using traditional, spatial, and object-based methods. *Wea. Forecasting*, **29**, 1451–1472, <https://doi.org/10.1175/WAF-D-13-00135.1>.
- Wong, M., and W. C. Skamarock, 2016: Spectral characteristics of convective-scale precipitation observations and forecasts. *Mon. Wea. Rev.*, **144**, 4183–4196, <https://doi.org/10.1175/MWR-D-16-0183.1>.
- Woodhams, B. J., C. E. Birch, J. H. Marsham, C. L. Bain, N. M. Roberts, and D. F. Boyd, 2018: What is the added value of a convection-permitting model for forecasting extreme rainfall over tropical East Africa? *Mon. Wea. Rev.*, **146**, 2757–2780, <https://doi.org/10.1175/MWR-D-17-0396.1>.
- Yan, H., and W. A. Gallus Jr., 2016: An evaluation of QPF from the WRF, NAM, and GFS models using multiple verification methods over a small domain. *Wea. Forecasting*, **31**, 1363–1379, <https://doi.org/10.1175/WAF-D-16-0020.1>.
- Yussouf, N., D. C. Dowell, L. J. Wicker, K. H. Knopfmeier, and D. M. Wheatley, 2015: Storm-scale data assimilation and ensemble forecasts for the 27 April 2011 severe weather outbreak in Alabama. *Mon. Wea. Rev.*, **143**, 3044–3066, <https://doi.org/10.1175/MWR-D-14-00268.1>.
- Zhang, F., C. Snyder, and R. Rotunno, 2003: Effects of moist convection on mesoscale predictability. *J. Atmos. Sci.*, **60**, 1173–1185, [https://doi.org/10.1175/1520-0469\(2003\)060<1173:EOMCOM>2.0.CO;2](https://doi.org/10.1175/1520-0469(2003)060<1173:EOMCOM>2.0.CO;2).
- Zhou, X., Y. Zhu, D. Hou, and D. Kleist, 2016: A comparison of perturbations from an ensemble transform and an ensemble Kalman filter for the NCEP Global Ensemble Forecast System. *Wea. Forecasting*, **31**, 2057–2074, <https://doi.org/10.1175/WAF-D-16-0109.1>.
- , —, —, Y. Luo, J. Peng, and D. Wobus, 2017: Performance of the new NCEP Global Ensemble Forecast System in a parallel experiment. *Wea. Forecasting*, **32**, 1989–2004, <https://doi.org/10.1175/WAF-D-17-0023.1>.