Assessing the Ensemble Predictability of Precipitation Forecasts for the January 2015 and 2016 East Coast Winter Storms

STEVEN J. GREYBUSH AND SETH SASLO

Department of Meteorology and Atmospheric Science, The Pennsylvania State University, University Park, Pennsylvania

RICHARD GRUMM

National Weather Service, State College, Pennsylvania

(Manuscript received 25 August 2016, in final form 9 January 2017)

ABSTRACT

The ensemble predictability of the January 2015 and 2016 East Coast winter storms is assessed, with model precipitation forecasts verified against observational datasets. Skill scores and reliability diagrams indicate that the large ensemble spread produced by operational forecasts was warranted given the actual forecast errors imposed by practical predictability limits. For the 2015 storm, uncertainties along the western edge's sharp precipitation gradient are linked to position errors of the coastal low, which are traced to the positioning of the preceding 500-hPa wave pattern using the ensemble sensitivity technique. Predictability horizon diagrams indicate the forecast lead time in terms of initial detection, emergence of a signal, and convergence of solutions for an event. For the 2016 storm, the synoptic setup was detected at least 6 days in advance by global ensembles, whereas the predictability of mesoscale features is limited to hours. Convection-permitting WRF ensemble forecasts downscaled from the GEFS resolve mesoscale snowbands and demonstrate sensitivity to synoptic and mesoscale ensemble perturbations, as evidenced by changes in location and timing. Several perturbation techniques are compared, with stochastic techniques [the stochastic kinetic energy backscatter scheme (SKEBS) and stochastically perturbed parameterization tendency (SPPT)] and multiphysics configurations improving performance of both the ensemble mean and spread over the baseline initial conditions/boundary conditions (IC/BC) perturbation run. This study demonstrates the importance of ensembles and convective-allowing models for forecasting and decision support for east coast winter storms.

1. Introduction

The east coast winter storms (ECWSs; Hirsch et al. 2001) of 26–27 January 2015 and 22–24 January 2016 delivered substantial impacts to the mid-Atlantic and northeastern United States. According to the Northeast Snowfall Impact Scale [NESIS; a regional snowfall index that ranks snowstorms as a function of area affected by the storm, the amount of snow, and the population living in the area impacted by the storm; Kocin and Uccellini (2004a)] the 25–28 January 2015 storm was ranked category 2 "*significant*," whereas the 22–24 January 2016 ranks fourth on the list with a category 4 "*crippling*" description. Notable were the forecast challenges, particularly near tight precipitation gradients at the northern and western edges of the storm where large ensemble spread occurred. In 2015, deterministic guidance indicated more than 2 ft of snow for New York City, New York, leading to its shutdown. A public outcry ensued when only 24.9 cm (9.8 in.) fell at Central Park, despite 63.2 cm (24.9 in.) occurring just 60 km to the east at Islip, New York, on Long Island, and 62.0 cm (24.4 in.) burying Boston, Massachusetts. Examination of operational ensemble forecasts [e.g., the Global Ensemble Forecast System (GEFS)], however, indicated a confident forecast for Boston, but large uncertainty in precipitation amounts for New York City. In 2016, ensembles confidently indicated a significant precipitation event in the Washington, D.C., metro area more than 4 days in advance, a forecast that successfully verified. However, until 24 h before the storm New York City appeared to be south of the main snow area, yet it received a record 69.9 cm (27.5 in.) at Central Park.

Corresponding author e-mail: Steven Greybush, sjg213@psu.edu

DOI: 10.1175/WAF-D-16-0153.1

While considerable progress has been made in identifying the patterns associated with (Miller 1946; Kocin and Uccellini 2004b; Root et al. 2007), understanding physical mechanisms responsible for (Sanders and Bosart 1985; Brennan and Lackmann 2005; Zhang et al. 2007; Ganetis and Colle 2015; Kumjian and Lombardo 2017), and modeling and predicting ECWSs (Evans and Jurewicz 2009; Novak et al. 2006, 2008), their practical predictability remains limited (Charles and Colle 2009a,b). Practical predictability is described as the ability to predict based on the procedures (e.g., observations, models, and assimilation systems) currently available (Melhauser and Zhang 2012). A particular challenge of these storms is the mesoscale nature of snowfall patterns, including banding structures that can result in snowfall rates of up to 7 cm h^{-1} on scales of tens of kilometers (Nicosia and Grumm 1999; Novak and Colle 2012). These are not adequately resolved in operational ensemble prediction systems, but require convection-permitting numerical weather prediction (NWP).

Lorenz (1963) demonstrated deterministic chaos: a sensitive dependence on initial conditions, where the present determines the future, but the approximate present does not approximately determine the future (as he later describes). Even with a perfect model, initial condition errors grow with time, resulting in intrinsic limits to predictability; imperfect models, observations, and assimilation schemes result in further limitations to the practical predictability of meteorological phenomena. Zhang et al. (2002) explored such predictability issues for the 2000 "*surprise*" snowstorm; errors can propagate upscale from convection, through gravity waves, and onto baroclinic instabilities (Zhang et al. 2007). Ensemble forecasting systems (Tracton and Kalnay 1993) are designed to address chaos by sampling plausible initial conditions (given observation and forecast uncertainties) and demonstrating an envelope of potential solutions. Today, ensemble generation is linked to the techniques of ensemble (Evensen 1994) and hybrid (Wang et al. 2013) data assimilation, with the goal of creating an ensemble with a calibrated spread that adequately describes forecast confidence. This is particularly important for quantitative precipitation forecasts (QPFs) for ECWSs, where decision-makers need to assess risk and take action (Novak et al. 2014).

Probabilistic precipitation forecasts from ensemble prediction systems are of limited value without understanding of their reliability. Siddique et al. (2015) demonstrated that a 100% probability forecast of precipitation from the raw ensemble (i.e., all members forecast precipitation exceeding a given threshold) verifies less than 70% of the time for precipitation in the mid-Atlantic region. Overall, ensemble forecasts for precipitation exhibit biases and tend to be underdispersive (Romine et al. 2014). Various techniques for assessing and correcting such deficiencies have been developed for coarser-resolution models (e.g., Raftery et al. 2005; Scheuerer and Hamill 2015), but have not yet been applied to convection-permitting ensembles whose performance characteristics may differ.

This paper assesses the performance of ensemble forecasts for the 25–28 January 2015 and 22–24 January 2016 winter storms using both operational ensemble forecasting systems and convection-permitting Weather Research and Forecasting (WRF) Model ensembles. The study explores the impact of initial conditions and ensemble design (i.e., modeling system, resolution, stochastic and physical perturbation method) on the practical predictability of these storms. Of particular interest is the link between the synoptic-scale evolution of the storm and precipitation patterns exhibited among ensemble members (explored using ensemble sensitivity techniques), time scales of predictability (displayed using a novel predictability horizon diagram), and an evaluation of the performance of the ensemble distributions (mean/median and spread) for precipitation. Section 2 describes these models and the precipitation datasets for comparison. Section 3 discusses the results in terms of ensemble skill scores, predictability horizons, sources of forecast uncertainty, and ensemble system design, and section 4 provides the conclusions.

2. Data and methods

a. *Operational ensemble models*

As of 2016, the National Centers for Environmental Prediction (NCEP) operates two primary ensemble forecast systems for weather prediction: the Global Ensemble Forecast System (GEFS), and the Short-Range Ensemble Forecast (SREF). The GEFS is a 21-member ensemble based upon the Global Forecast System (GFS) model. In winter 2015, the operational deterministic resolution was T574 (~25 km), and the ensembles were run at T254 (~55 km). By winter 2016, this had been increased to TL1534 (~13 km), with the ensembles at TL574 (~33 km). The GEFS uses initial condition diversity through ensemble transformation with rescaling (Wei et al. 2008) for 2015, and then hybrid–ensemble data assimilation (EnDA) perturbations in 2016 with a single dynamical core and set of physics packages. To enhance ensemble spread during the forecast phase, stochastic total tendency perturbations (STTPs) are used, which perturb the total model state tendency every 6 h using a random perturbation.

During the winter of 2015, the Short-Range Ensemble Forecasts (Stensrud et al. 2000) system featured 21

TABLE 1. Configurations of the ensemble model simulations used in this study. Note that the SREF uses several analyses for the initial conditions (indicated by asterisks). Select WRF experiments employ SKEBS, SPPT perturbations, both (STOC), or multiphysics (PHYS).

Event	Expt	IC–BC	Perturbations	Physics	Resolution (km)
2015	GEFS	GDAS	IC + STTP	Single	55
2015	SREF	GDAS*	IC + multicore	Multi	16
2016	GEFS	GDAS	IC + STTP	Single	33
2016	SREF	GDAS*	IC + multicore	Multi	16
2016	BWW	GEFS	IC only	Single	20
2016	NCAR	EnKF	IC only	Single	3
2016	WRF-GEFS	GEFS	IC only	Single	3
2016	WRF + SKEB	GFS	SKEBS only	Single	3
2016	WRF-GEFS + SKEB	GEFS	IC + SKEBS	Single	3
2016	WRF-GEFS + SPPT	GFES	IC + SPPT	Single	3
2016	WRF-GEFS + STOC	GEFS	IC + SKEBS + SPPT	Single	3
2016	WRF-GEFS + PHYS	GEFS	IC only	Multi	3

members at ~16 km horizontal resolution. Members were evenly divided (seven each) among the Nonhydrostatic Mesoscale Model (NMM), NMM-B, and Advanced Research version of WRF (ARW) dynamical cores. Initial conditions were created using three pairs of positive and negative perturbations generated by the bred vector technique (Toth and Kalnay 1993). By winter 2016, the SREF underwent a significant upgrade, with 26 members evenly divided among the NMM-B and ARW dynamical cores. The new SREF also includes diversity in initial conditions and model physics options (Du et al. 2015).

b. *Convection-permitting WRF ensemble simulations*

The need for a convection-permitting ensemble system has been identified as an important strategic priority by the UCAR Model Advisory Committee (UMAC) in their recent review of the NCEP production suite (Carr and Rood 2015), which involved stakeholders across the weather enterprise including the operational, research, and private-sector communities. Running models at convective scales (generally less than 4 km) removes the need for a convective parameterization, and allows for a more realistic depiction of vertical motion, cloud and precipitation features, and interaction of flow with topography. Currently, a 4-km North American Mesoscale Forecast System (NAM) nest and 3-km High-Resolution Rapid Refresh (HRRR; Benjamin et al. 2016) provide the state of the art for operational prediction at convective scales in the United States. However, a single deterministic run does not provide an indication of forecast confidence and, therefore, may be misleading in regard to which features can be relied upon, and what details are beyond the limit of predictability. Previous ensemble studies at convection-permitting resolutions for specific events have focused primarily on severe weather over the Great Plains (e.g., Stensrud et al. 2009) and tropical cyclones (e.g., Zhang et al. 2009). The National Center for Atmospheric Research (NCAR) WRF Data Assimilation Research Testbed (WRF-DART) ensemble and its website (Schwartz et al. 2015; http://ensemble.ucar.edu/), which started running in April 2015 in near–real time over the CONUS at 3 km once a day at 0000 UTC, are excellent resources for forecasters and researchers; however, this system was not yet operational during the 2015 case. This ensemble continuously cycles (every 6 h) a 15-km WRF ensemble Kalman filter (EnKF) single-physics ensemble of 50 members, then launches ten 3-km ensemble member forecasts at 0000 UTC each day (Schwartz et al. 2015). Model physics include the Thompson microphysics, Rapid Radiative Transfer Model for GCMs (RRTMG) radiation, Noah land surface model, and MYJ PBL schemes. For comparison, we include results from a special run of the NCAR ensemble for the 2016 case here, where 30 members were launched at 3-km resolution at 1200 UTC 22 January 2016.

The NSF-funded Big Weather web project (BWW; Maltzahn et al. 2016) has developed a multiuniversity ensemble system to test advances in cyberinfrastructure to enable easier creation, reproducibility, sharing, and analysis of large datasets. This ensemble employs a CONUS domain with 20-km grid spacing (approximately the resolution of the operational ensembles), and provides a WRF-based test bed for ensemble design. A special run of the BWW ensemble created by the authors uses 20 members with the 1200 UTC 22 January 2016 GEFS for initial and boundary conditions, and WRF physics options selected by the BWW team (similar to the NCAR ensemble, with the addition of the Kain–Fritsch cumulus scheme).

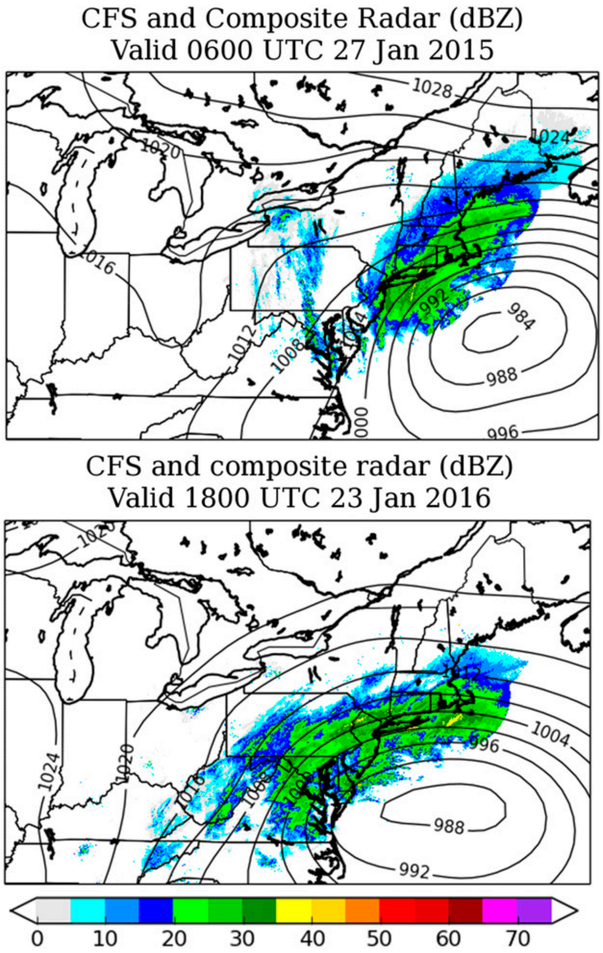
A convection-permitting (3-km resolution) ensemble forecast employing the WRF Model (Skamarock et al. 2008) was created by the authors by downscaling each member of the GEFS. A triply nested domain was used for the 2015 case, with resolutions of 27 km (outer domain covering the entire eastern United States), 9 km (middle domain covering the Carolinas to Maine), and 3 km (inner domain covering Virginia to New Hampshire's southern border). For the 2016 case, only the 9- and 3-km domains were needed because of the increased resolution of the NCEP models. WRF is initialized with GEFS data at 27 vertical levels, but WRF is run with 43 vertical levels. As each WRF member uses initial conditions (ICs) and boundary conditions (BCs) from a corresponding GEFS member, the GEFS data assimilation system provides the initial ensemble perturbations (ideally, toward the fastest-growing synoptic-scale instabilities), as well as flow-dependent boundary condition uncertainty to match the global model.

The baseline ensemble configuration was single core, single physics by design (in contrast to the SREF), so that the focus would be on initial condition uncertainties. The physics configuration is the Thompson two-moment microphysics (Thompson et al. 2008), the RRTM longwave (Mlawer et al. 1997) and Dudhia (1989) shortwave radiation schemes, the Eta surface layer (Janjić 1996, 2002), the Noah land surface model (Chen and Dudhia 2001), the MYJ boundary layer scheme (Janjić 1994), and Grell's cumulus [at 9- and 27-km domains only; Grell and Dévényi (2002)].

In special experiments corresponding to the 2016 case, additional ensemble configurations were explored. These made use of stochastic perturbations to model forecasts, including the stochastic kinetic energy backscatter scheme (SKEBS; Shutts 2005; Berner et al. 2011) and the stochastically perturbed parameterization tendency (SPPT; Palmer et al. 2009; Berner et al. 2015). Another configuration uses multiple combinations of cloud microphysics and boundary layer schemes. Table 1 summarizes the configurations of these runs.

c. Observation datasets

Model forecasts are verified using several observation products. In this study, we have focused on the verification of precipitation. Models do not necessarily provide a conversion from liquid equivalent to snow depth, and observations have demonstrated broad variability in this ratio even within the same storm. For example, for the 2015 storm Boston had a snow/liquid ratio (SLR) of 23.2:1, whereas Central Park showed 11.7:1. We note difficulties with rain gauge data [such as the Automated Surface Observing System (ASOS)], including undermeasurement bias during windy conditions (e.g., Doesken and Robinson 2009). We focus on snow water equivalent (SWE) for QPF produced by models as well as for observation products; the challenges associated with SLR (e.g., Roebber et al. 2003; Baxter et al. 2005) do not need to be addressed here.


CFS and Composite Radar (dBZ)
Valid 0600 UTC 27 Jan 2015

CFS and composite radar (dBZ)
Valid 1800 UTC 23 Jan 2016

FIG. 1. Surface pressure analyses from the Climate Forecast System Reanalysis (CFSR; hPa; black contours) and observed composite radar reflectivity (dBZ; shaded) during the height of the (top) Jan 2015 and (bottom) Jan 2016 snowstorms.

Liquid equivalent precipitation from ASOS and Cooperative Observer Program (COOP) stations provide the best in situ measurements, but are somewhat limited in spatial coverage. These can be enhanced by community efforts such as CoCoRAHS; however, this network does not enforce the collection of reports at a uniform time of day. National Weather Service (NWS) public information statements provide a valuable resource of snowfall amounts measured by trained SKYWARN spotters as well as members of the public, with fine spatial resolution; however, liquid equivalents are not provided.

To resolve the detailed structure of snowfall accumulation patterns, a spatially dense dataset is desired. A mosaic of composite reflectivity radar imagery provides snapshots of the evolution of the storm structure, and can be compared with model-simulated reflectivity

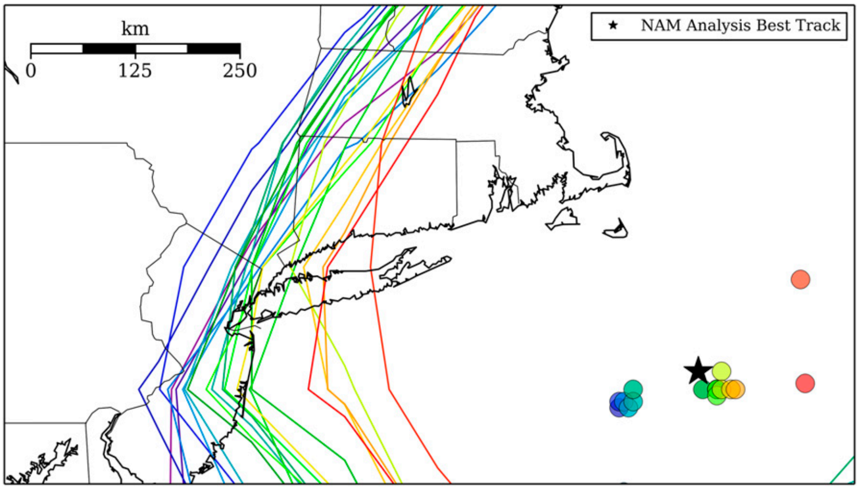


FIG. 2. Locations of storm centers as estimated from minimum sea level pressure from GEFS ensemble forecasts initialized at 1200 UTC 26 Jan 2015 and valid at 1200 UTC 27 Jan 2015. Location of minimum pressure from the verifying NAM analysis is shown as a black star. Points are colored according to their longitudinal distance from the analysis, with purple being farthest west and red farthest east. Contours indicate the westernmost extent of the 25.4-mm storm total precipitation threshold, colored by its respective GEFS member.

(using the WRF version 3.7 forward operator, which maps microphysical state variables to reflectivity). Raw radar precipitation estimates are subject to known errors and biases, due to limitations in elevation angle, interference from topography, range folding, beam attenuation, and precipitation-type uncertainties (Nelson et al. 2016). The NCEP stage IV precipitation product (Seo 1998) calibrates gridded, radar-estimated precipitation with surface stations, and provides a best-guess observational product for use in this study.

3. Results

a. *Synoptic overview of the snowstorms*

The blizzard of 2015 was a significant east coast winter storm (DeGaetano et al. 2002) that brought areas of snow to portions of the mid-Atlantic region and the northeastern United States. The heaviest snow fell over eastern Long Island northward into eastern New England and southeastern Maine. The storm evolved as a northern stream short-wave trough [clipper; e.g., Hutchinson (1995)] moved into the mid-Atlantic region then off the east coast, providing favorable upper-level vorticity advection (downstream of the trough) and divergence (in the left-exit region of an upper-level jet streak) patterns. This system merged with a southern stream trough, which led to rapid cyclogenesis (e.g., Gaza and Bosart 1990), aided by diabatic heat release from condensation. The resulting cyclone and radar echoes are shown at 0600 UTC 27 January 2015 (Fig. 1a). A broad area of precipitation, mainly snow, was present on the western side of the cyclone from western Long Island northeastward into southeastern Maine. Within the broader precipitation shield a more intense mesoscale snowband was present over eastern Long Island. This band shifted northward into southeastern New England between 0600 and 1200 UTC (not shown). The sharp western gradient of heavy snow for this storm was associated with the position of the attendant coastal low pressure area.

Using the GEFS ensemble, the linkages between the track of the coastal low, the western extent of precipitation, and the upper-air fields a few days prior are demonstrated. In Fig. 2, the position for the surface low in each ensemble member is indicated (the clustering of positions is an artifact of the discrete, coarse nature of the GEFS grid). A coastal low is a characteristic ‘‘fingerprint’’ of east coast winter storms (Root et al. 2007). The western edge of the 25.4-mm precipitation contour is indicated with matching colors. Note the nearly direct correspondence between the position of the low and the westward extent of the heavy snow. The middle cluster of ensemble members, which agrees with the actual storm position, correctly places this threshold right through New York City. Given the uncertainty in the initial conditions at 1200 UTC 26 January, there would be no way to determine a priori which solution would verify. One can trace the track errors to uncertainties in the upper-air pattern at model initialization times

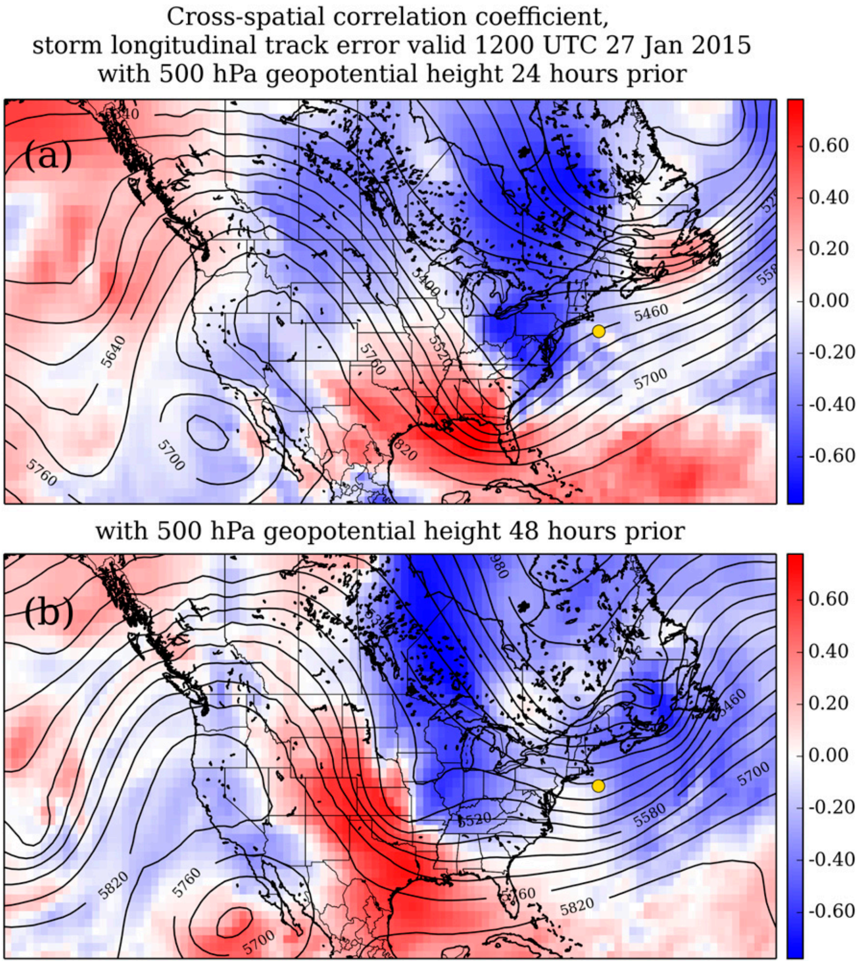

Cross-spatial correlation coefficient,
storm longitudinal track error valid 1200 UTC 27 Jan 2015
with 500 hPa geopotential height 24 hours prior
(a)
5460
5700
5640
5700
5760
5400
5520
5760
5820
0.60
0.40
0.20
0.00
-0.20
-0.40
-0.60
with 500 hPa geopotential height 48 hours prior
(b)
5460
5700
5580
5520
5820
5760
5820
5700
0.60
0.40
0.20
0.00
-0.20
-0.40
-0.60

FIG. 3. Ensemble sensitivity as demonstrated by correlations between GEFS track error (storm position at verification time indicated as yellow dot; east = positive error, west = negative error) as correlated with time-lagged 500-hPa heights: (a) 24 h and (b) 48 h prior.

(Fig. 3). Ensemble sensitivity analyses (Bishop et al. 2001; Ancell and Hakim 2007; Torn and Hakim 2008; Zheng et al. 2013; Ota et al. 2013) reveal the correlations between ensemble perturbations of a scalar forecast metric (here, track error) and a model field (here, 500-hPa heights). Here, track errors are defined as the distances between low pressure centers in the GEFS forecasts compared to the NAM verifying analysis in the east–west direction (irrespective of latitude), with a forecast location east of the observed location being a positive number. A positive correlation between track error and the 500-hPa height field (red colors) indicates regions where higher heights encourage a track farther out to sea, and lower heights encourage a track closer to the coast. Therefore, a deeper (or slower) 500-hPa trough over Alabama (at $T - 24$ h) or Kansas (at $T - 48$ h) may have brought the storm westward and with it heavier snow to New York City. We further illustrate the relationship between storm track and precipitation in section 3b.

The blizzard of 2016 produced record to near-record snows from the Washington, D.C., area to New York City based on National Weather Service observations. The heaviest snow was observed across northeastern West Virginia and southwestern Pennsylvania northeastward to the New York metropolitan area. Snowfall totals in the New York metropolitan area ranged from 59.4 cm (23.4 in.) at Islip to 69.9 cm (27.5 in.) in Central Park, and 77.5 cm (30.5 in.) at John F. Kennedy International Airport (JFK) in Queens County. Reports of over 75 cm were observed in southern Pennsylvania. The blizzard of 2016 developed as a southern stream wave moved up the east coast. Regions of snow initiated in the strong easterly flow north of the deepening cyclone as it moved across North Carolina and over the western Atlantic. At 1200 UTC 23 January 2016 the cyclone was

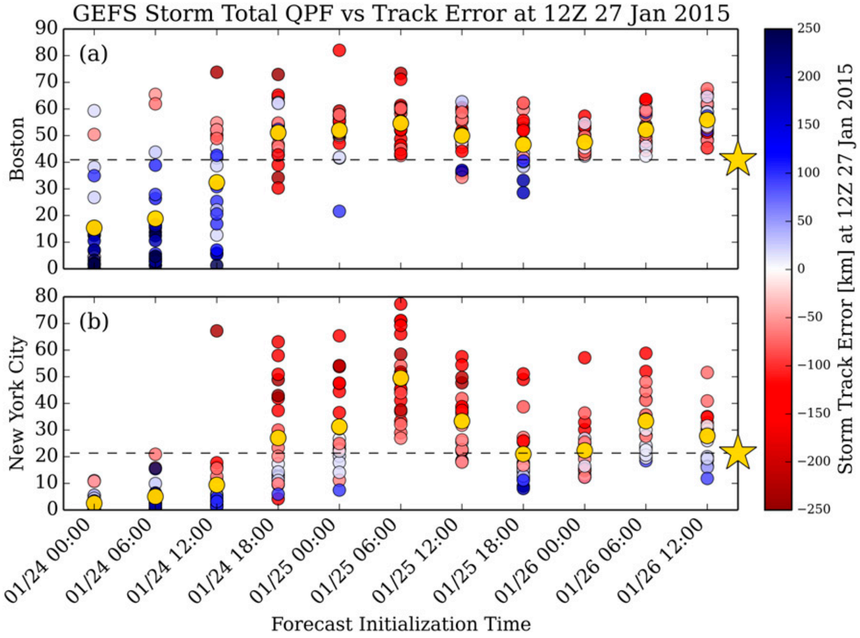


FIG. 4. Predictability horizon diagram for liquid equivalent precipitation (mm) at (a) Boston and (b) NYC. GEFS ensemble storm total precipitation forecasts ending at 1200 UTC 28 Jan 2015 are compared to storm track errors evaluated at 1200 UTC 27 Jan. Track error is defined as the longitudinal distance of the storm center from the verifying NAM analysis, as shown in Fig. 2. Negative values (red) indicate a westward displacement and positive (blue) eastward. The horizontal axis indicates the time of forecast initialization. The dashed lines and stars on the rightmost vertical axes indicate the observed liquid equivalent precipitation. Yellow dots are ensemble mean values; other dots are individual ensemble member forecasts.

east of the Delmarva region with a broad precipitation shield extending from western Maryland to southern New York State. Embedded within this region of snow were several strong mesoscale bands of heavier snow. By 1900 UTC (Fig. 1b), the cyclone shifted to the east and the heavier snow shifted into northeastern Pennsylvania through southern New England.

b. Predictability horizons

Consider an idealized schematic for ensemble predictability horizons (Figs. 4 and 5 demonstrate actual examples, which are elucidated later). The horizontal axis represents the lead time prior to an extreme event, which occurs at the far right of the figure. The vertical axis denotes the ensemble forecasts of an important event parameter (e.g., track of a low pressure area, amount of snowfall, etc.) initialized at a particular time relative to the event (x axis). All forecasts, however, are *valid* at the same time (the event time). Therefore, the diagram shows how forecasts evolve as the event approaches. The middle curve (in Fig. 5) is the ensemble mean/median, whereas the bottom/top curves represent the ensemble spread (one standard deviation, as in Fig. 5), minimum/maximum, or percentiles and define the envelope of potential solutions. The distance between the curves indicates the ensemble spread. The first derivatives (slopes) of the curve indicate forecast trends, whereas the second derivatives indicate forecast jumpiness/consistency.

This type of diagram can illustrate key time scales that define the predictability horizons for an event. At the far left of the figure, probabilities for the extreme event are expected to be near climatology. A first critical time scale is identified when *initial detection* for the event occurs: a few ensemble members indicate the possibility for an extreme event, but the likelihood of the event, as well as its specific details, remain unclear. There may be considerable run-to-run inconsistencies at this stage. A second critical time scale is when the *emergence of a signal* occurs: a significant subset of the ensemble (e.g., >50%) agrees that an extreme event may occur, and therefore this signal is indicated in the ensemble mean or median. However, the ensemble spread remains large, indicating that several scenarios are still plausible. A third critical time scale takes place when a *convergence of solutions* occurs around a single outcome. While alternative scenarios are still possible, they are less likely to occur. The ensemble spread has become

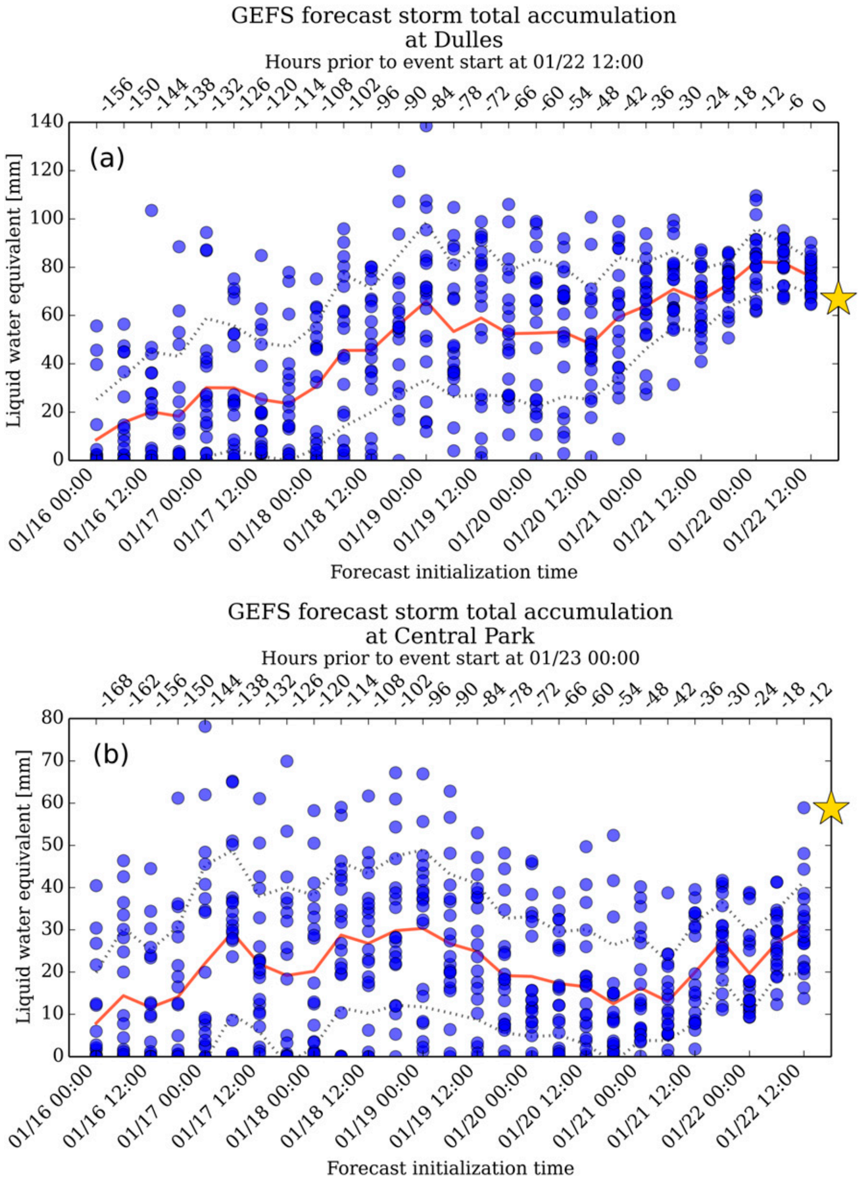


FIG. 5. Ensemble predictability horizon diagram for GEFS ensemble storm total precipitation forecasts ending at 1200 UTC 25 Jan 2016 for (a) DC and (b) NYC. Blue dots indicate individual GEFS ensemble member forecasts, the red line indicates the ensemble mean forecast, and gray dashed lines indicate one standard deviation of the ensemble. The bottom horizontal axis shows the forecast initialization date and time, while the top axis shows this time as hours prior to the start of the event (distinct for each location). Predictability can be assessed in three stages: initial detection, emergence of a signal, and convergence of solutions.

small, and the forecast confidence subsequently becomes high.

Figure 4 depicts a predictability horizon diagram for precipitation in New York City and Boston for the January 2015 storm. For Boston, initial detection of the event has occurred by 0000 UTC 24 January 2015, with emergence of a signal for significant snow by 1800 UTC 24 January. By 0000 UTC 26 January, the ensemble is confident in >40 mm of SWE, which verifies (albeit on the lower end of the ensemble envelope). For New York City, the initial detection and emergence of a signal are delayed in time relative to Boston, and a convergence of solutions never occurs. This indicates that the uncertainty in the initial conditions and model error do not allow storm track scenarios to be ruled out even 12–24 h prior to the storm. In this diagram, the

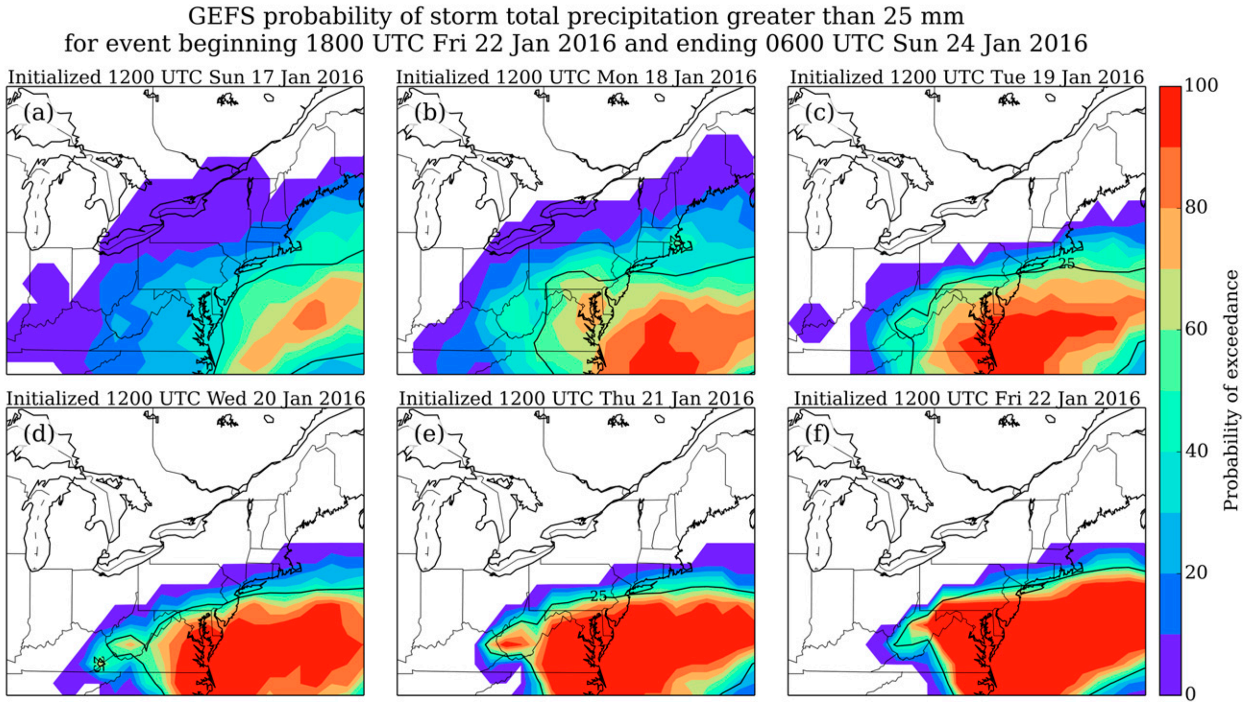

GEFS probability of storm total precipitation greater than 25 mm
for event beginning 1800 UTC Fri 22 Jan 2016 and ending 0600 UTC Sun 24 Jan 2016
Initialized 1200 UTC Sun 17 Jan 2016
(a)
Initialized 1200 UTC Mon 18 Jan 2016
(b)
Initialized 1200 UTC Tue 19 Jan 2016
(c)
25
Initialized 1200 UTC Wed 20 Jan 2016
(d)
Initialized 1200 UTC Thu 21 Jan 2016
(e)
25
Initialized 1200 UTC Fri 22 Jan 2016
(f)
Probability of exceedance
100
80
60
40
20
0

FIG. 6. Ensemble probability plots for 25-mm liquid equivalent. Forecasts are from the GEFS initialized at 1200 UTC (a) 17, (b) 18, (c) 19, (d) 20, (e) 21, and (f) 22 Jan 2016. As the event approaches, the ensemble demonstrates higher confidence for significant precipitation, particularly in the DC metro area.

ensemble member forecasts are colored by (east–west) storm track error (see Fig. 2 for a visual depiction of low pressure error locations). Errors in precipitation are strongly correlated with the east–west displacement of the storm: note that higher than average QPFs for New York are nearly always accompanied by red dots (storm tracking closer to the coast), whereas the ensemble members that verify closest to the observations tend to show little to no storm track error (white shading).

We also examine the predictability of the 22–24 January 2016 storm. Predictability horizons (Fig. 5) for the Washington, D.C. (DC), area were long: more than 6 days for initial detection and 4 days for emergence of a signal, with a convergence of solutions taking place ~36 h prior to the onset of the event. The situation was more complicated for New York; whereas initial detection was also early, the ensemble never fully converged on a solution. In this event, unlike the 2015 event, the verifying observation was near the highest ensemble member. Overall, this storm had a significantly longer practical predictability horizon than the 2015 storm. Figure 5 also shows several examples of a bimodal precipitation distribution, where the ensemble mean is actually a not especially likely scenario.

Figure 6 depicts the evolution of probability maps for 25 mm of precipitation as a function of GEFS forecast initialization time. Initial detection of the possibility of an extreme event occurred by 17 January, with greater than 50% confidence for the DC area appearing by 18 January. By 1200 UTC 21 January, the forecast reached 90% confidence for the DC metro area, whereas the New York City area remained near 50%. The northern and western gradients remained a considerable forecast challenge.

These results illustrate that the question "*how far in advance was the storm predictable?*" does not always have a simple answer. Each aspect of the storm can be traced through the stages of initial detection, emergence of a signal, and convergence of solutions (if it occurs). Synoptic-scale features of a storm have longer predictability horizons (e.g., the formation of an intense low pressure area off the eastern seaboard), whereas details (exact location of the low, locations of mesoscale snowbands and the northwest edge of precipitation) take longer to appear.

To be a useful source of forecast confidence, an ensemble system must be reliable: over a significant number of cases, an event forecasted with X% probability must occur approximately X% of the time. Reliability diagrams are typically developed over many cases to provide a large statistical sample. However, forecasters may wish to know the conditional reliability of an

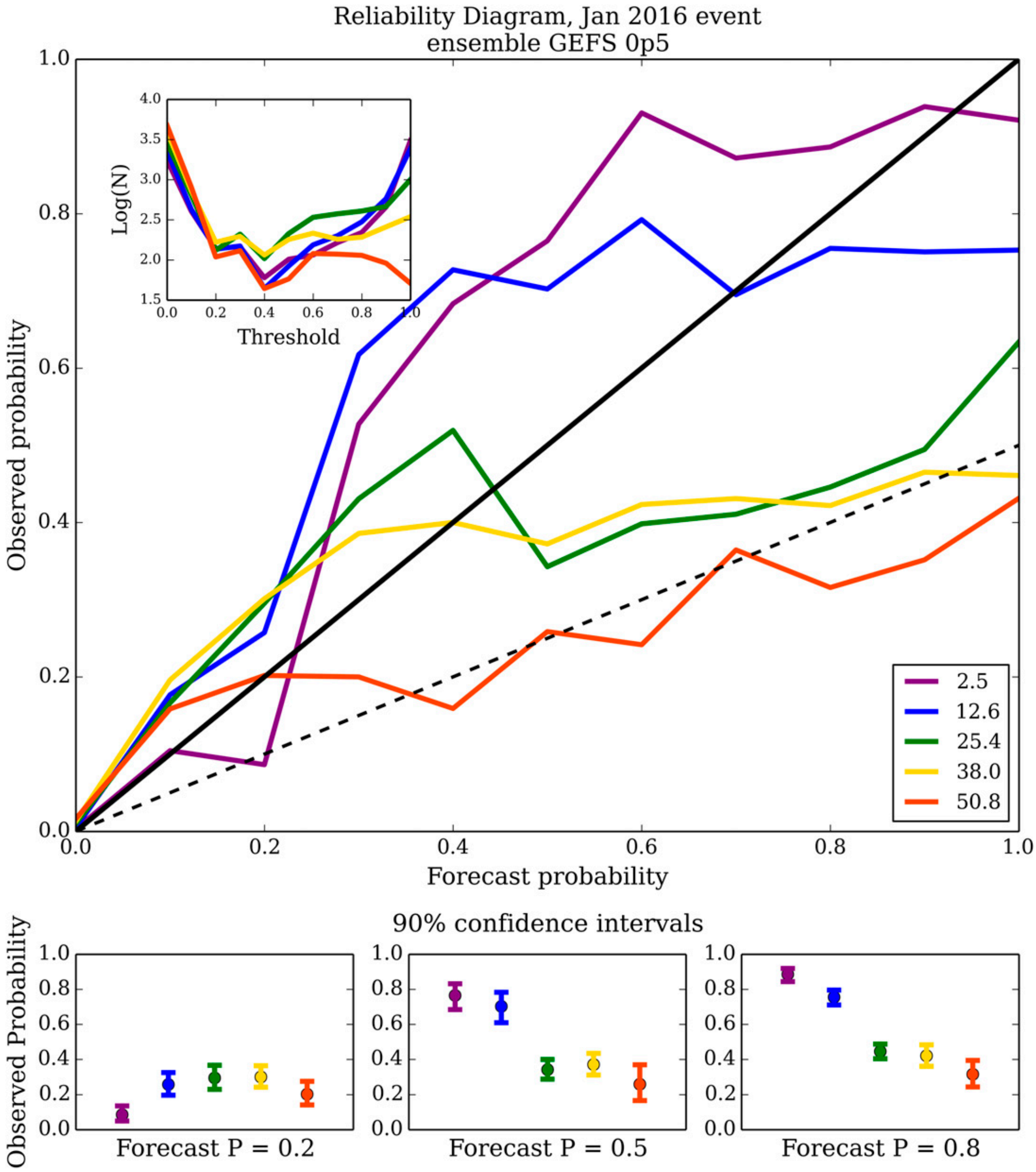


FIG. 7. Reliability diagram illustrating the performance of the storm total precipitation forecasts from the GEFS ensemble for the Jan 2016 event. Colored lines indicate the ensemble forecast probability compared to the observed frequency for five precipitation thresholds (mm). The black solid line is the line of perfect reliability, while the dashed line represents the line of "no skill" (as in Wilks 2011). Forecasts are compiled from six initializations of GEFS prior to the start of the event, at approximately 0000 UTC 23 Jan. The 0.5° data are sampled. Observations are taken from the U.S. Global Historical Climatology Network (GHCN) database. Inset shows the logarithm of the number of observations in each probability bin.

ensemble for a specific weather regime: for example, if an underforecast bias suddenly becomes an overforecast bias for intense snowfalls. Therefore, we have created a reliability diagram for ensemble probabilities of precipitation thresholds for the 2016 event (Fig. 7), gathering samples spatially as well as temporally (for different lead times, rather than independent events). The GEFS is interpolated to COOP (GHCN) precipitation locations, resulting in a sample size of thousands of points (inset). We note that the effective degrees of freedom, however, are considerably smaller as these points are not independent because of spatial and temporal correlations. The 90% confidence intervals for the observed probabilities were computed

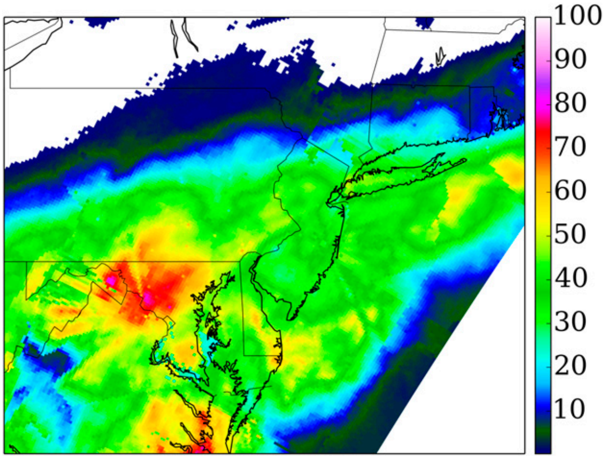


FIG. 8. Stage IV precipitation estimate of 72-h storm total accumulated liquid equivalent precipitation (mm), beginning at 1200 UTC 22 Jan and ending at 1200 UTC 25 Jan 2016.

using the Jeffreys interval for binomial distributions (Brown et al. 2001), shown for forecast probability thresholds of 0.2, 0.5, and 0.8 in Fig. 7. For small thresholds (2.5 and 12.6 mm) there is an underforecasting bias; for example, a 50% probability of precipitation actually occurs 70% of the time. For larger thresholds (25.4 and 38 mm), an overforecasting bias occurs, and the ensemble is overconfident. For example, an ensemble forecast of 25 mm occurring in all members actually verified only in 60% of cases. There is little skill in the 50-mm predictions. The reliability of an ensemble forecast can be improved through ensemble design (perturbation methods, as discussed in section 3c) as well as postprocessing methods.

c. Evaluating ensemble design

This section focuses on the impact of ensemble system design (e.g., resolution, single–multiphysics, perturbation method) on both forecast ensemble mean and spread. Figure 8 depicts the observed storm total precipitation from stage IV estimates for the January 2016 ECWS. Observations indicate the most precipitation occurred north and west of the DC metro area toward the Appalachian Mountains, then extending just west of the Interstate Highway 95 (I-95) corridor toward New York City. There is a very tight precipitation gradient through central Pennsylvania, with a distance of only ~50 km separating locations that received 45 cm of snow from those reporting only a few centimeters.

Next, precipitation forecasts from the operational GEFS, SREF, and several WRF ensemble systems are

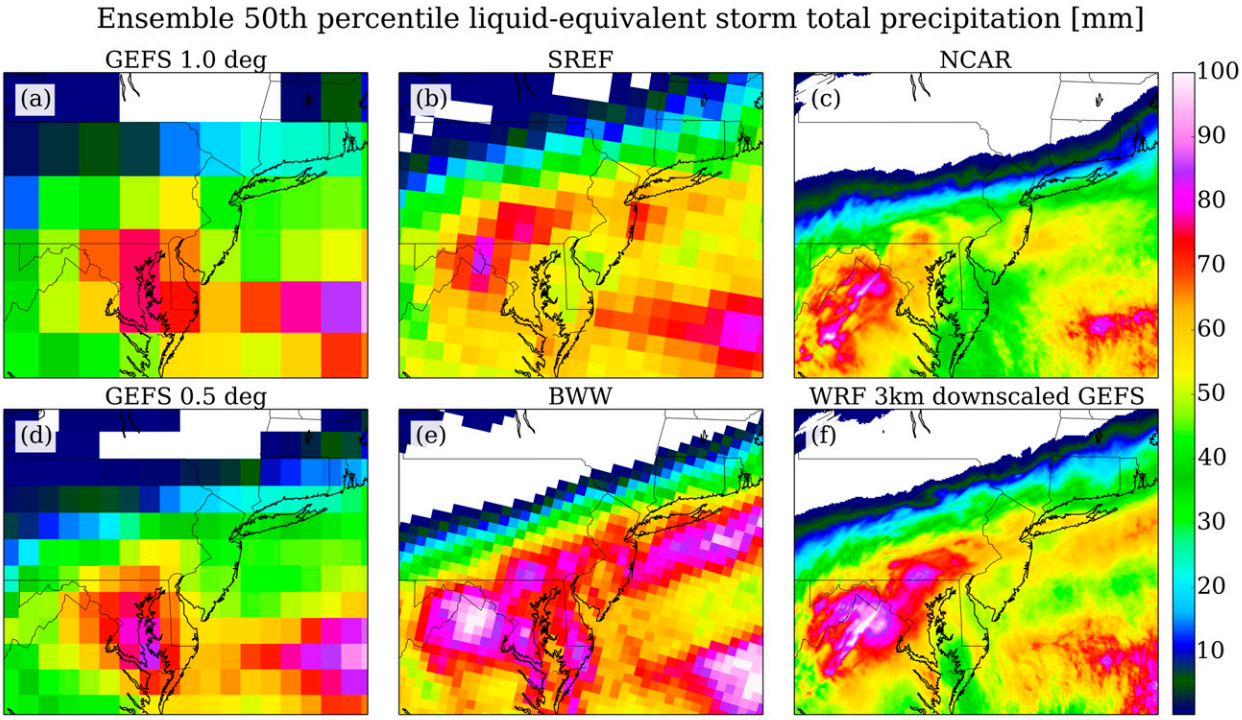


FIG. 9. Ensemble 50th percentile storm total liquid equivalent precipitation forecast by the (a) GEFS 1.0°, (b), SREF, (c) NCAR 3-km ensemble, (d) GEFS 0.5°, (e) BWW ensemble, and (f) WRF 3 km downscaled from the GEFS. The 50th percentile indicates that at a grid point, 50% of ensemble members forecast an amount less than or equal to the amount shown. Sizes of the squares indicate the resolution of the product.

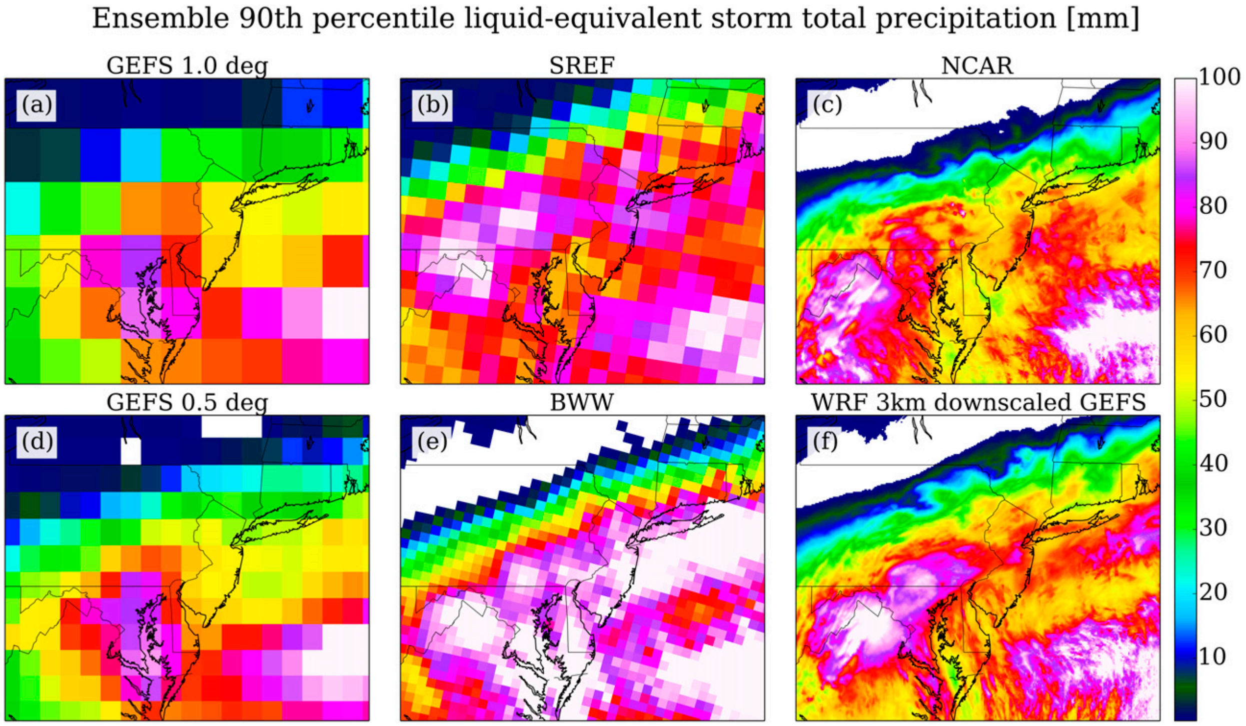


FIG. 10. As in Fig. 9, but for the 90th percentile.

compared. To encourage probabilistic rather than deterministic thinking, the ensemble 50th percentile (median; Fig. 9), 90th percentile (Fig. 10), and 10th percentile (Fig. 11) are displayed. The coarse resolution of 1.0° GEFS (panel a in Figs. 9–11) compared to 0.5° GEFS (panel d in Figs. 9–11) illustrates the importance of using the highest resolution available when verifying precipitation forecasts. All model forecasts appear to overpredict the maximum precipitation over land (~70 mm in the stage IV product), with 80–100-mm simulated QPFs. The placement of the heaviest precipitation was too far south and east with the GEFS; it was correctly shifted north and west in the WRF downscaled versions. While the SREF (panel b in Figs. 9–11) correctly identified heavier precipitation northwest of the DC area and into southeastern Pennsylvania, the precipitation expanded too far northward. Switching from the GEFS to a WRF-based ensemble (panels c, d, and f in Figs. 9–11) resulted in a generally larger QPFs over land. As this was true for both the convective-parameterized BWW (panel e in Figs. 9–11) as well as the convective-permitting run (panel f in Figs. 9–11), both resolution and model physics/dynamics played a role in modifying precipitation fields. The WRF 3 km downscaled from GEFS corrected a spurious precipitation maxima south of Long Island, and better resolved the impact of topography on snowfall amounts over the Appalachians, as did the NCAR ensemble (panel d in Figs. 9–11), albeit with a more southerly track for the storm.

When comparing differences among ensemble means of the various 3-km WRF ensembles (Fig. 12), it is important to focus on large spatial areas of systematic differences. The blotchy patterns found off the mid-Atlantic coast are noise due to heavy convective precipitation being located at slightly different locations in each ensemble member. Despite the noisy plots, there are some regions with important signals. For example, the addition of stochastic perturbations tended to reduce the ensemble mean QPFs by 5–10 mm in the region north and west of the I-95 corridor. One might expect that positive and negative stochastic perturbations should tend to cancel each other out in the ensemble mean field. This systematic shift can have several potential causes: sampling error (from a limited ensemble size and, therefore, perturbations may not exactly cancel) or nonlinear effects (the ensemble mean may be close to ‘‘optimal’’ for heavy precipitation in this region, and any disruption to this scenario leads to smaller total precipitation amounts). This difference pattern was even more pronounced for the multiphysics ensemble (indicating nonlinear sensitivity to the selection of PBL and microphysics schemes) and the NCAR ensemble mean. The combination of Thompson two-moment

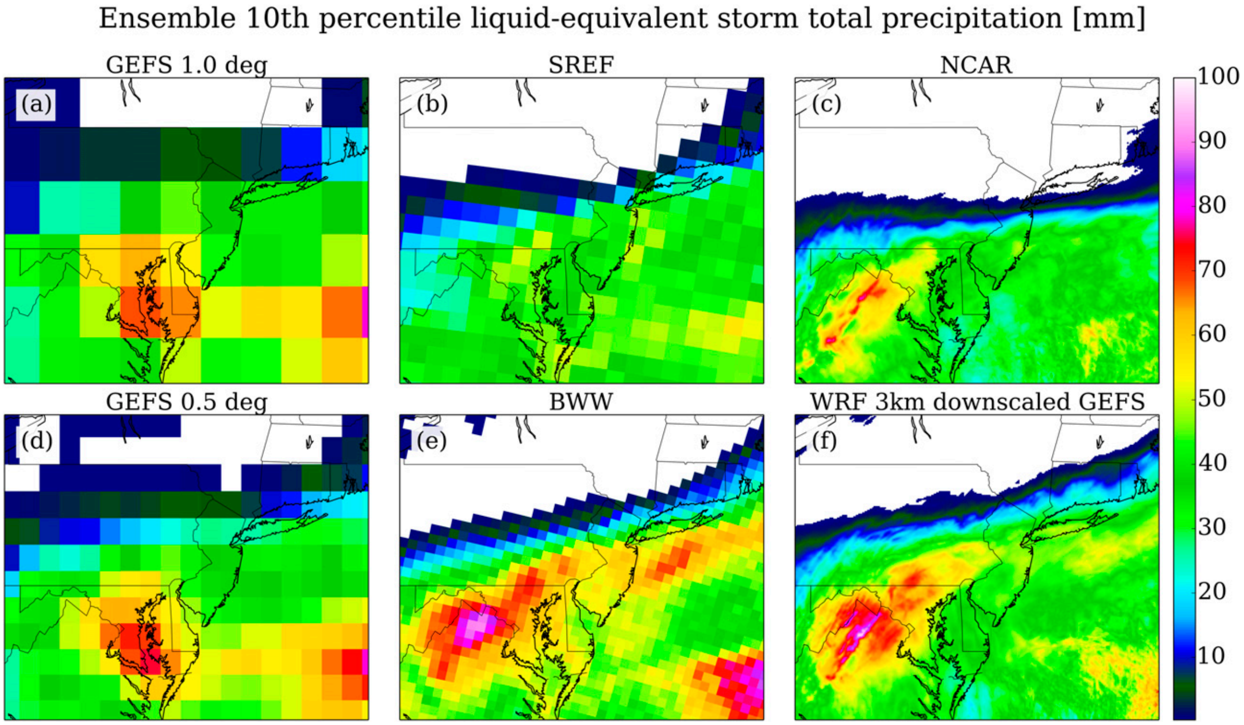


FIG. 11. As in Fig. 9, but for the 10th percentile.

microphysics and the MYJ PBL may therefore produce greater QPFs than other physics options employed in the multiphysics ensemble.

The impact of the ensemble perturbation method on spread (standard deviation) is illustrated in Fig. 13. Among the operational models, the GEFS showed the smallest spread. The SREF, which varies dynamical cores, ICs, and physics, has a very large spread in the northern extent of the precipitation as several members brought heavy precipitation much farther north. It is interesting to note that simply downscaling from the GEFS provides some additional spread, particularly in the offshore convective region, and in the higher elevations of the Appalachians. Zhang et al. (2007) posed a mechanism by which small-scale convective instabilities project onto unbalanced (which radiate away as gravity waves) and balanced synoptic-scale components, which can excite baroclinic instabilities. The 3-km simulations explicitly permit convection, which is particularly active in the warm sector offshore. Mesoscale processes not explicitly resolved by the global model, such as differences in updraft locations, interactions with topography, and the location and timing of mesoscale snowbands, may result in greater ensemble spread. In Fig. 14, the baseline configuration is WRF 3 km downscaled from the GEFS, with spread shown as a difference from the spread in this run. Again, the signal here is with systematic differences over land, whereas noisy differences offshore should be ignored. The addition of SKEBS and SPPT increased the spread over southern Pennsylvania and Maryland, in the regions of heavy snow. The multiphysics WRF ensemble (the most "SREF like" of the experiments) had larger spread in the region of heaviest precipitation, but smaller spread across the fringe areas. This is indicative of feedbacks between convective precipitation and changes in latent heating (in physical tendencies or explicitly in alternate microphysics and PBL schemes). The NCAR ensemble, which cycles its DA independently from the GEFS (other than BCs), shows a larger area of ensemble spread (>10 mm) through the I-78 corridor of southeastern Pennsylvania (Harrisburg to Allentown); as this ensemble is single physics, this is an indication that its initial conditions exhibited a larger spread than the operational GEFS. Therefore, differences in spread among ensemble systems are a consequence of differing initial spread (e.g., NCAR versus GEFS), model characteristics (GEFS versus WRF), resolution (GEFS versus downscaled; generally larger spread for convective permitting), physics (single versus multiple schemes; generally larger spread for multiphysics), and stochastic perturbation type.

To further explore ensemble precipitation forecasts, box-and-whisker diagrams were generated for nine

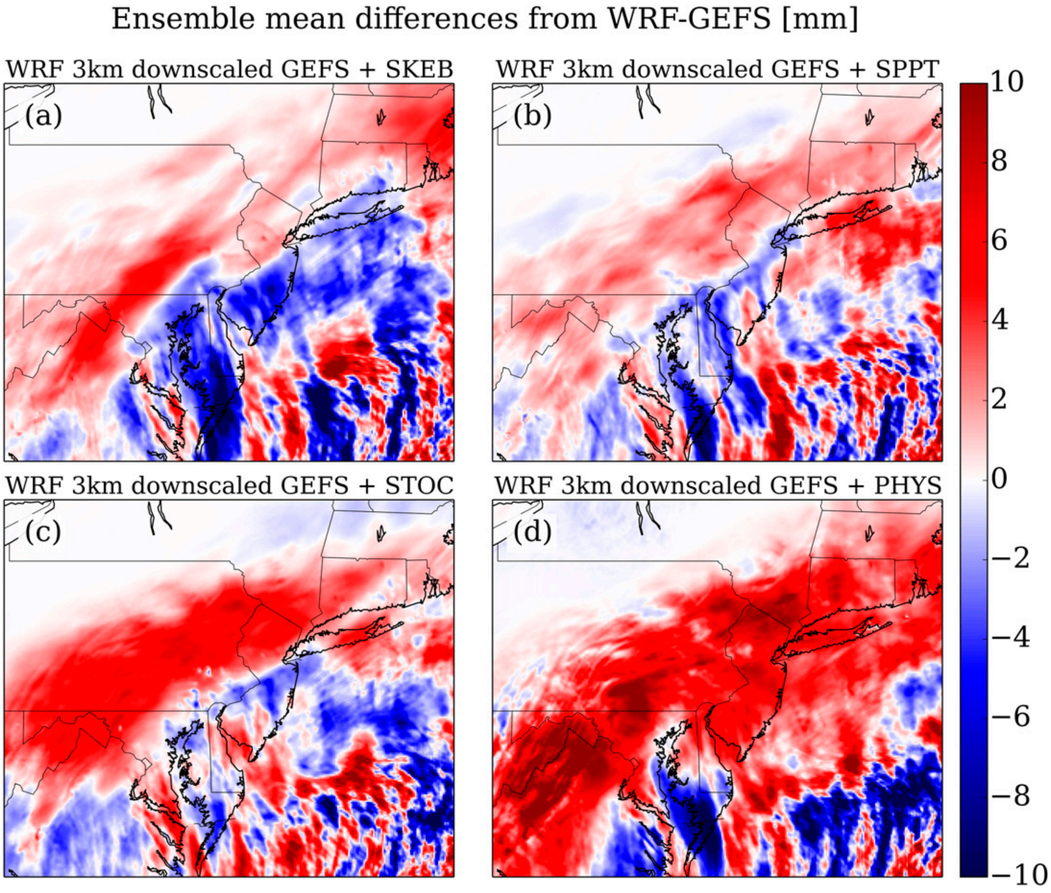

Ensemble mean differences from WRF-GEFS [mm]
WRF 3km downscaled GEFS + SKEB
(a)
WRF 3km downscaled GEFS + SPPT
(b)
WRF 3km downscaled GEFS + STOC
(c)
WRF 3km downscaled GEFS + PHYS
(d)
10
8
6
4
2
0
−2
−4
−6
−8
−10

FIG. 12. Differences in ensemble mean precipitation, shown as the WRF-GEFS run minus (a) WRF-GEFS + SKEB, (b) WRF-GEFS + SPPT, (c) WRF-GEFS + STOC, and (d) WRF-GEFS + PHYS. Red indicates the original WRF-GEFS ensemble had higher precipitation values.

locations (Sterling, Virginia; Altoona, Pennsylvania; State College, Pennsylvania; Harrisburg, Pennsylvania; Bethlehem, Pennsylvania; Central Park; Upton, New York; Storrs, Connecticut; and Taunton, Massachusetts); we feature three representative plots in Fig. 15. First, we examine the performance of the operational systems. At Sterling, all forecast systems produced a significant SWE event, which verified (not shown). Several locations, such as Altoona, State College (Fig. 15c), Storrs, and Taunton (Fig. 15b), were located along the northern edge of the verifying QPE shield. Harrisburg, Bethlehem, Upton, and Central Park (Fig. 15a) all verified with sufficient SWE for heavy snow but were north of the axis of highest forecast SWE.

At Central Park (Fig. 15a), the difficult forecasting question was that of a routine, or instead, a high-impact winter storm; the latter was observed. The operational GEFS implied a low-end snow event and the SREF indicated potential for a high-impact heavy snow event with nearly 60 mm of SWE. The GEFS grossly underforecast the SWE at Central Park and the SREF, skewed toward the wet side, better matched the observed precipitation. The downscaled ensemble members provided significantly more snow and (generally) smaller spread than the original GEFS, providing useful guidance.

For Taunton (Fig. 15b), the SREF forecasted a high probability for QPF supporting large SWE with very large ensemble spread, while the GEFS forecast a low-end snow event. In Taunton the observed SWE was close to the median forecast in the GEFS and near the lower limits of the SREF ensemble. The experimental guidance generally was biased toward the wetter SREF solution, with larger spread than the GEFS. From a forecast perspective, the issue was differentiating between a high-impact winter storm or a low-end snow event; the low-end snow event was observed.

At State College (Fig. 15c), and with the exception of the SREF, most of the guidance had a sharp northern QPF shield and produced QPF values implying a

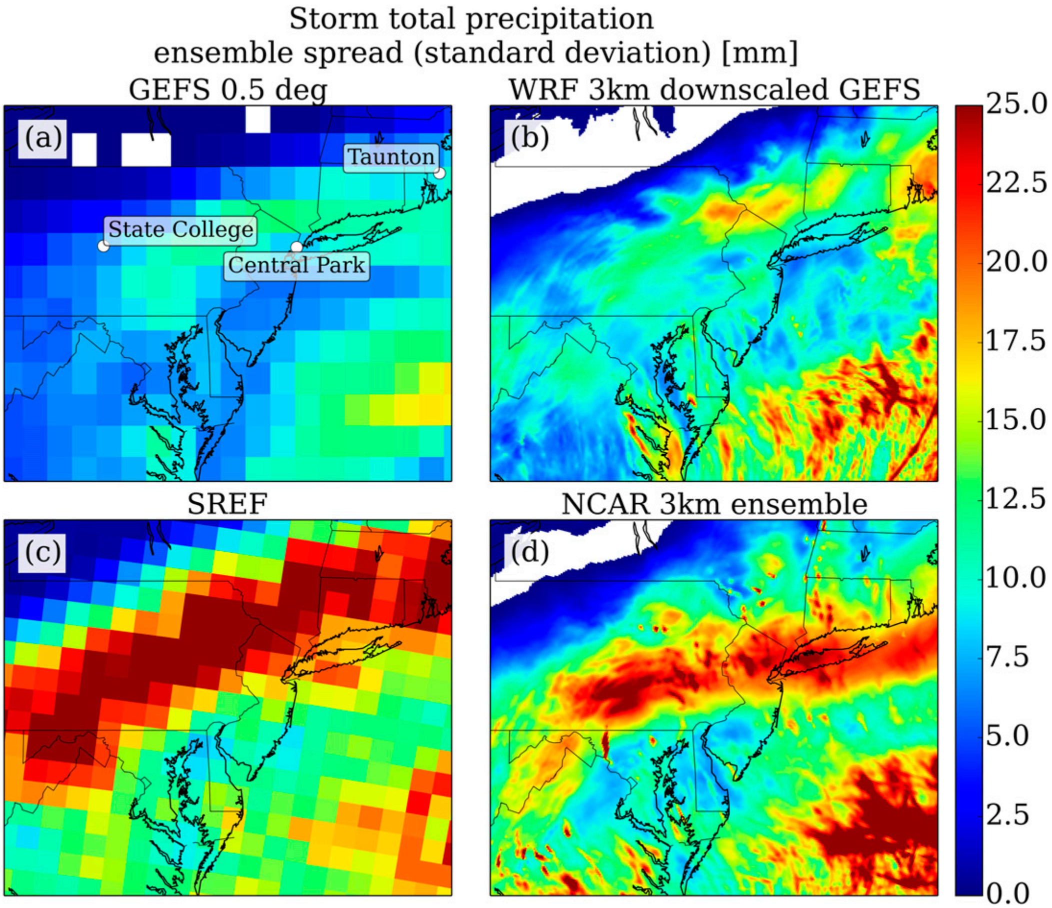


FIG. 13. Ensemble spread (standard deviation from the ensemble mean) of precipitation forecasted by the (a) GEFS 0.5°, (b) WRF 3 km downscaled from the GEFS, (c) SREF, and (d) NCAR 3-km ensemble.

low-end snow event. The SREF forecasts indicated a high probability of a significant snow event, but with large uncertainty values. Operationally, the difference between the SREF and GEFS was significant in distinguishing between a winter storm warning or a winter weather advisory; the downscaled GEFS members supported a low-end winter storm with much smaller uncertainty than the SREF. A low-end heavy snow event was observed in State College though snow amounts there varied from 25 cm in southern areas to under 5 cm a few miles north of town. In in the end, State College had a low-end winter storm with warning criteria snow.

An optimal design for a convective-permitting ensemble system produces precipitation forecasts with both an accurate ensemble mean and an appropriate ensemble spread. Simply switching to WRF (BWW ensemble) and downscaling the GEFS to 3 km dramatically increased the QPFs, with the deterministic run matching closely the observed snow water equivalent at Central Park and State College. Overall, the downscaled GEFS ensembles produced both greater precipitation and larger spread than the original GEFS. The SREF ensemble mean verified well for New York; however, it tended to push the heavy precipitation too far north in other regions. The NCAR ensemble had extremely large spread for New York City, with the mean significantly underforecasting the QPF. As coarse-resolution, single-physics ensembles can be underdispersive, we examined the role of stochastic perturbations and multiphysics in increasing the ensemble spread at the convective-allowing scales. SPPT was more effective than SKEBS at increasing (and, in this case, improving) ensemble spread, with the multiphysics being the most effective. We also explored changing the spatial scale and amplitude of SKEBS and SPPT from their default configuration (not shown), but did not see a large sensitivity in ensemble mean and spread for total precipitation. Ensemble spread produced stochastically (with SKEBS and SPPT) could qualitatively compare to

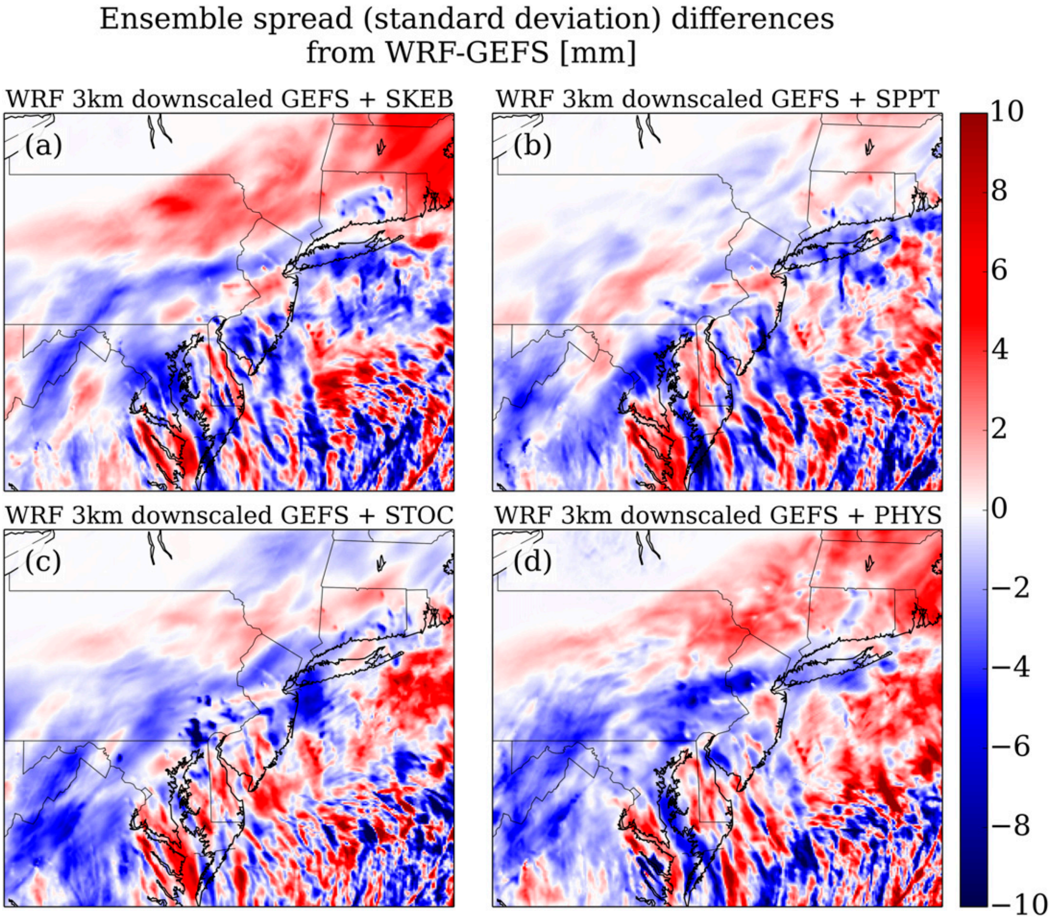

Ensemble spread (standard deviation) differences from WRF-GEFS [mm]
WRF 3km downscaled GEFS + SKEB
(a)
WRF 3km downscaled GEFS + SPPT
(b)
WRF 3km downscaled GEFS + STOC
(c)
WRF 3km downscaled GEFS + PHYS
(d)
10
8
6
4
2
0
−2
−4
−6
−8
−10

FIG. 14. Differences in ensemble spread of precipitation, shown as the WRF-GEFS run minus (a) WRF-GEFS + SKEB, (b) WRF-GEFS + SPPT, (c) WRF-GEFS + STOC, and (d) WRF-GEFS + PHYS.

that of a multiphysics ensemble. SREF and NCAR ensembles showed even larger spread, but with an overforecasting bias in the SREF and an underforecasting bias in the NCAR ensemble.

Table 2 quantifies spatial average performance statistics for the various ensembles. The Brier (1950) score (similar to the mean squared error but for binary events) is evaluated for the ensembles with respect to various precipitation thresholds. The continuous ranked probability score (CRPS; Matheson and Winkler 1976; Hersbach 2000) is the integral of the Brier score over all possible thresholds and is useful for verifying ensemble performance; it reduces to the mean-squared error for deterministic forecasts (lower values indicating better performance). Hersbach (2000) decomposes the CRPS into three components: reliability, resolution, and uncertainty. We focus on the reliability component (related to the rank histogram), where a low value indicates a well-calibrated ensemble where probabilities "mean what they say" in that forecasted probabilities match actual event probabilities (Wilks 2011). The uncertainty term is entirely a function of the sample climatology; greater resolution describes the ability of the forecast PDF to discern events with greater sharpness than climatology. The GEFS, NCAR ensemble, and 3-km WRF ensembles demonstrated clear superiority in all metrics to the SREF for the 2016 storm; the GEFS scores better than the SREF for the 2015 case as well. Skill scores for the GEFS and convection-permitting ensembles are of a similar magnitude. As Mass et al. (2002) indicate, jumping to higher resolution (from 12 to 4 km in their case) produces more realistic detail and structure for weather features, but does not necessarily improve traditional verification scores as mesoscale features are relatively underconstrained by data and their position errors can be amplified. The probabilistic information from ensembles has clear value beyond a deterministic, high-resolution run for evaluating forecast confidence. Among the 3-km downscaled runs, the ensemble simulations downscaled from GEFS clearly

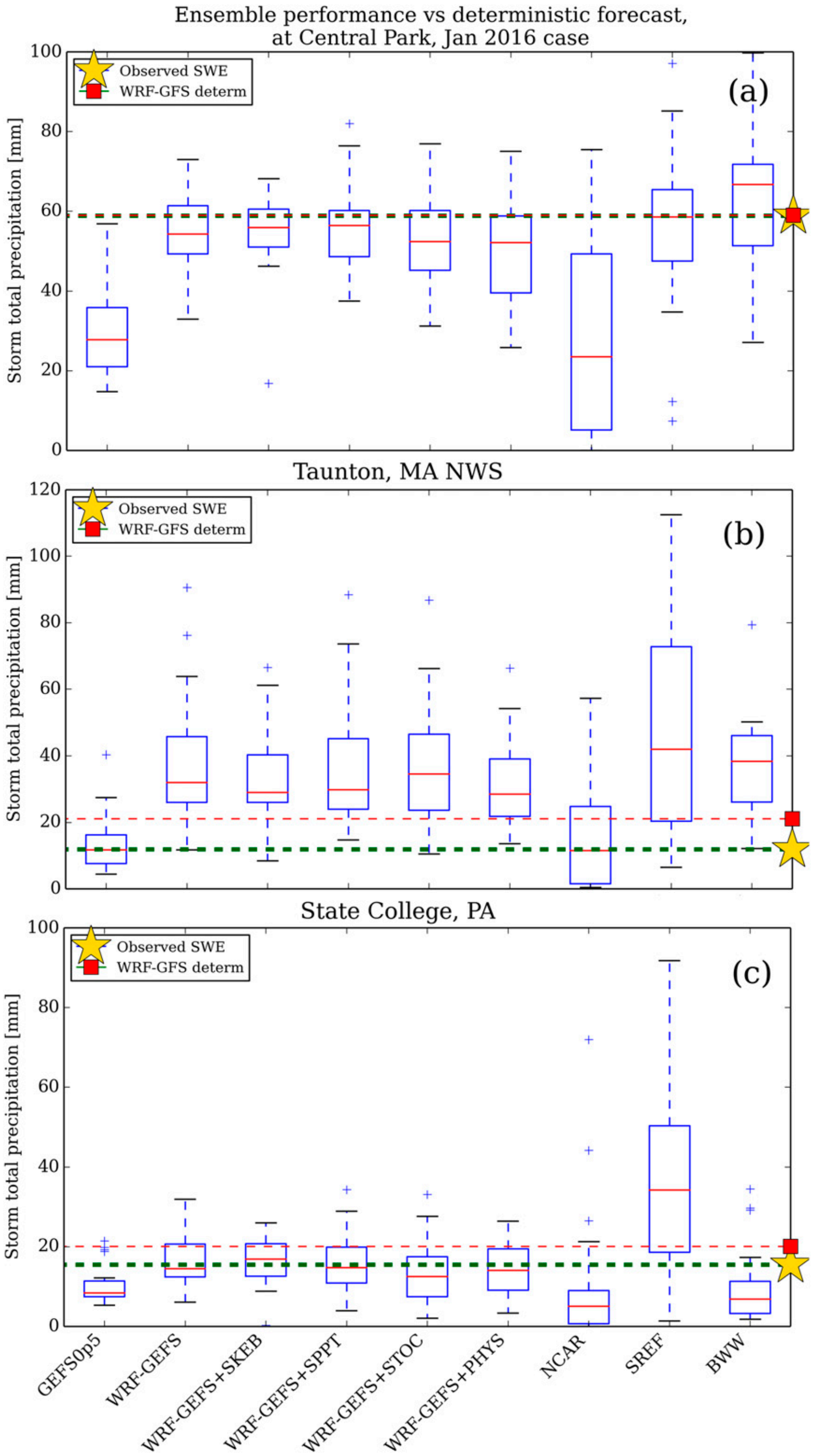


FIG. 15. Box-and-whisker plots showing storm total precipitation forecasts at (a) NYC, (b) Taunton, and (c) State College for all ensembles, for the event ending at 1200 UTC 25 Jan 2016. Red line segments indicate the median value, the box extends from the bottom to the top quartile of the ensemble forecasts, and whiskers extend to 1.5 times the interquartile range. Outliers are marked by small blue plus signs. All ensembles are initialized at 1200 UTC 22 Jan, with the exception of SREF, which is initialized at 0900 UTC. Red box indicates the forecast from a 3-km resolution WRF forecast initialized using the deterministic GFS. Gold star and horizontal green dashed line show the observed liquid equivalent.

TABLE 2. Performance of model precipitation forecasts assessed through CRPS, the reliability component of CRPS, and the Brier score for three precipitation thresholds (12.6, 25.4, and 38 mm) for the Jan 2016 storm (forecasts initialized at 1200 UTC 22 Jan 2016; top rows) and the Jan 2015 storm (forecasts initialized at 1200 UTC 26 Jan 2015; bottom rows).

Expt	Mean CRPS	Reliability	12.6-mm BS	25.4-mm BS	38-mm BS
GEFS 1.0 (2016)	6.238	3.858	0.136	0.089	0.073
GEFS 0.5 (2016)	6.149	3.798	0.129	0.087	0.072
SREF (2016)	9.309	6.039	0.160	0.201	0.123
BWW	8.663	5.991	0.133	0.147	0.115
NCAR3 km	6.910	3.208	0.134	0.118	0.089
WRF-GEFS	6.638	3.937	0.133	0.105	0.078
WRF + SKEB	7.060	5.295	0.140	0.110	0.084
WRF-GEFS + SKEB	6.483	3.667	0.131	0.101	0.078
WRF-GEFS + SPPT	6.480	3.714	0.131	0.101	0.078
WRF-GEFS + STOC	6.494	3.461	0.134	0.104	0.080
WRF-GEFS + PHYS	6.088	3.116	0.127	0.100	0.071
Deterministic 3 km	9.311		0.154	0.129	0.101
GEFS 1.0 (2015)	7.746	6.413	0.185	0.168	0.125
SREF (2015)	10.671	8.848	0.261	0.222	0.166

improved upon the score of the deterministic forecast downscaled from the GFS. The addition of stochastic perturbations resulted in a slight improvement in CRPS and reliability compared to the baseline downscaled ensemble. It was important to perturb both the IC/BC at the forecast initiation as well as stochastically during the forecast phase: WRF-GEFS + SKEB beat WRF + SKEB (no IC/BC perturbations) and WRF-GEFS (no stochastic perturbations). Among various perturbation options, the multiphysics ensemble had the best (lowest) scores in all categories. Using both SKEB and SPPT resulted in a similar CRPS, but improved reliability.

d. *Predictability of mesoscale features*

A convection-permitting ensemble can provide insights into the predictability of mesoscale features in ECWSs. The distribution of snowfall totals depends not only on synoptic-scale factors, such as the position of the primary low pressure area (Fig. 2), but also moist convective processes, which have considerably shorter predictability horizons. Zhang et al. (2003) illustrated how moist convective errors can project onto baroclinic instabilities that impact the synoptic-scale evolution of the system.

Figure 16 compares ''paintball'' plots of the simulated radar reflectivity at 1900 UTC 23 January 2016 (a 31-h forecast), a time when intense snowbands were occurring. Figure 16a provides a sense of the intrinsic predictability: differences in ensemble member forecasts are due to initial condition uncertainty only, not model error. The only way to gain confidence in the location of the snowbands is to further refine the GEFS initial conditions (albeit these are controlled by the data assimilation system, observations, and prior forecasts). Figure 16b shows how the ensemble members (each member, tied to a specific GEFS member for the initial conditions) change their forecasts in response to stochastic perturbations. These perturbations can be thought of as a proxy for model error, particularly processes inadequately resolved in the model physics that feed back to the dynamics. As the color coding matches between panels (i.e., the ensemble member initiated from GEFS member 1 is shaded the same), we can see the impact of perturbations in physics on the location of mesoscale snowbands (also see Figs. 16c,d). For example, probabilities for precipitation in southeastern Pennsylvania are greater in the GEFS ensemble compared to the SPPT ensemble (Figs. 16e,f). This further illustrates the necessity of a probabilistic or ensemble approach to the prediction of these features, given the demonstration of sensitivity to both the initial conditions and model error.

4. Conclusions

The ensemble predictability of precipitation for two intense east coast winter storms in January 2015 and 2016 is examined. Both storms provided an excellent case study for probabilistic forecasts, with high-confidence forecasts indicated for Boston in the 2015 storm and for Washington, D.C., in the 2016 storm, but with a large ensemble spread in snowfall amounts for New York City (NYC) in both storms. QPFs from ensemble forecasting systems are validated against in situ COOP observations and the stage IV multisensor precipitation product. This large spread was warranted, as 2015 verified near the lower end of the envelope, and 2016 at the upper end (a new record total); the official deterministic forecasts were poor, but the ensembles contained the verifying solution within their envelope.

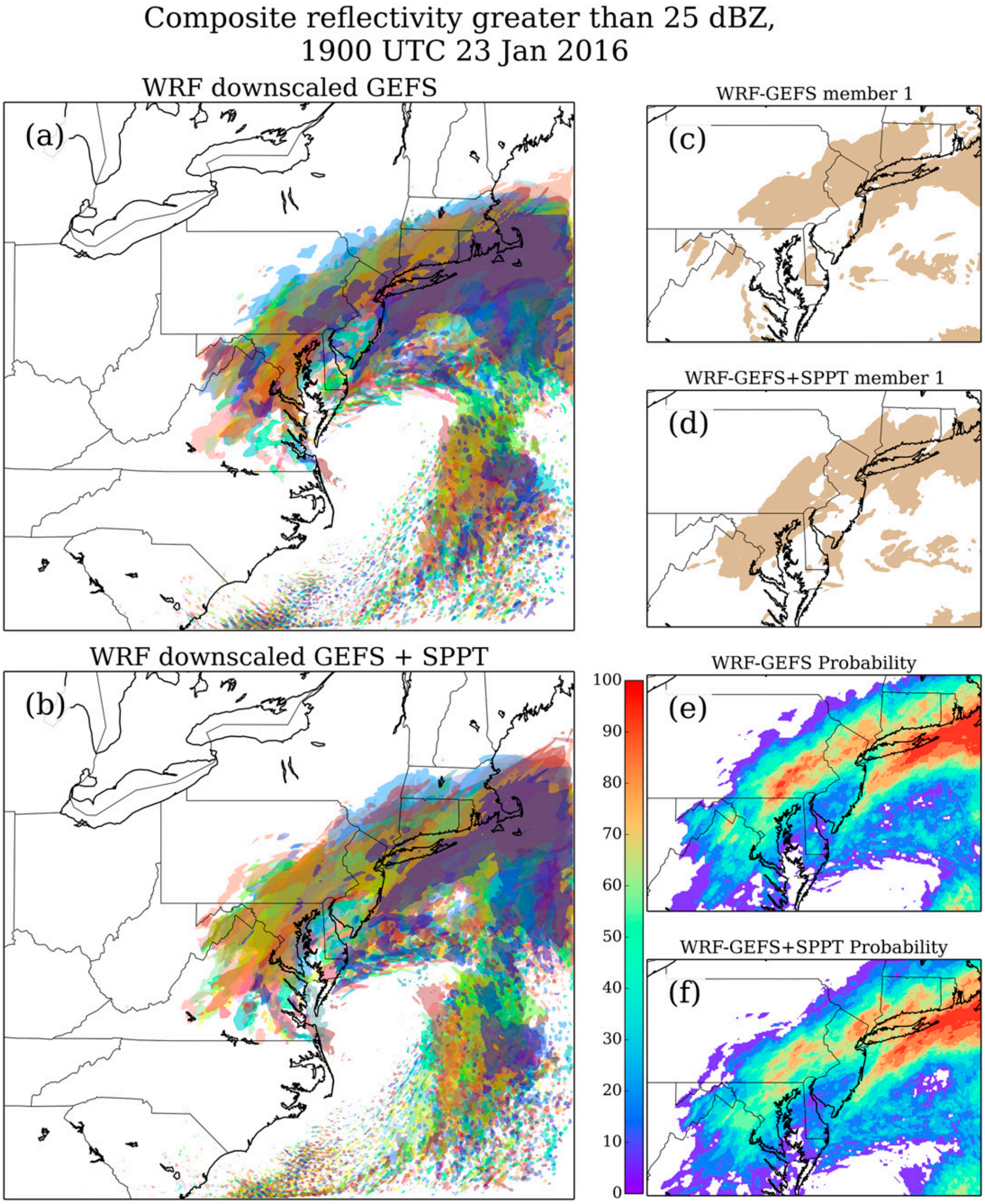


FIG. 16. Paintball plots for composite reflectivity for the (a) WRF 3 km initialized from the GEFS and (b) WRF 3 km initialized from the GEFS with SPPT valid at 1900 UTC 23 Jan 2016. Each ensemble member is assigned a color, and regions exceeding a threshold of 25 dBZ receive a translucent fill. The 25-dBZ threshold was selected to highlight areas of significant precipitation, such as mesoscale snowbands. (c),(d) Ensemble member 1 is highlighted for GEFS and GEFS + SPPT, respectively, illustrating the impact of stochastic physics perturbations on the intensity of this individual reflectivity field. (e),(f) The ensemble probability of exceeding 25 dBZ at all grid points for GEFS and GEFS + SPPT, respectively, showing the impact of stochastic physics perturbations on the ensemble as a whole.

Indeed, reliability diagrams (compiled for the 2016 storm at all locations and multiple forecast lead times) indicated that forecasts were still overconfident–underdispersive, requiring even greater ensemble spread to match the forecast errors. For the 2015 storm, the forecasted snow for the NYC area was strongly related to small east–west position errors in the attendant coastal surface low, as a tight western gradient of the precipitation developed. Using the ensemble sensitivity technique, these position errors could be traced to the timing/position of antecedent 500-hPa troughs 1–2 days earlier; an improvement in initial conditions (observations/DA) might have better constrained this trough position and, subsequently, increased practical predictability for NYC.

Predictability horizon diagrams can indicate the forecast lead time in terms of (i) initial detection, (ii) emergence of a signal, and (iii) convergence of solutions for an event. These differ considerably by storm and by feature. For example, for the 2016 storm initial detection (at the synoptic scale) occurred at least 6 days in advance (considerably more than the 2015 storm), whereas mesoscale predictability of snowbands did not converge even 24 h prior to the event. Zhang et al. (2003) demonstrated that small-scale perturbations can grow rapidly as a result of convective instabilities, whereas perturbations projected onto baroclinic instabilities grow more slowly but may reach larger amplitudes. The increase in spatial coverage of high-confidence forecasts for heavy snow can be tracked in ensemble probability plots, a potential forecast tool for the real-time evaluation of major precipitation events. It is important that these forecasts are well calibrated for extreme events such as east coast winter storms. The predictability diagrams also provide good visualization of uncertainty for both operational forecasters and students learning about uncertainty in weather forecasting.

Convective-scale ensembles promise to be the future of operational NWP, with the ability to explicitly resolve mesoscale features such as squall lines, banding features in winter cyclones, and terrain interactions. Therefore, exploration of an optimal ensemble design, development of a high-resolution regional assimilation system, use of innovative graphical displays, and validation of performance of such a system for high-impact weather such as ECWSs and extreme precipitation events in the highly populated northeastern United States is an important prerequisite for forecasting and decision support capabilities. Several sets of ensemble forecasts at 3-km resolution downscaled from the GEFS using WRF were run for the 2016 storm using various perturbation methods: IC/BC only, SKEBS, SPPT, both, and multiphysics. All perturbation methods improved upon the control case, with the multiphysics scoring best for CRPS and reliability with respect to precipitation observations. Finally, paintball plots provided some insights into both the intrinsic and practical predictability (by comparing IC only and stochastically perturbed ensembles) of precipitation features such as mesoscale snowbands; slight changes in initial conditions or small perturbations during the forecast phase could lead to considerable position and timing differences for these bands. When using the next generation of convection-permitting NWP to predict east coast winter storms and communicating forecasts to end users, it is important to keep in mind predictability limits and provide adequate notions of forecast confidence.

Acknowledgments. Portions of this work were funded by NSF Award 1450488. The authors thank Craig Schwartz for providing access to a special run of the NCAR ensemble, thank three anonymous peer reviewers for helpful comments on this manuscript, and acknowledge NOAA for access to operational model output and observation products. The authors are grateful to the Penn State Department of Meteorology and Atmospheric Science and the Institute for CyberScience for computational resources used in this study.

REFERENCES

Ancell, B., and G. J. Hakim, 2007: Comparing adjoint- and ensemble-sensitivity analysis with applications to observation targeting. *Mon. Wea. Rev.*, **135**, 4117–4134, doi:10.1175/2007MWR1904.1.

Baxter, M. A., C. E. Graves, and J. T. Moore, 2005: A climatology of snow-to-liquid ratio for the contiguous United States. *Wea. Forecasting*, **20**, 729–744, doi:10.1175/WAF856.1.

Benjamin, S. G., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Rea. Rev.*, **144**, 1669–1694, doi:10.1175/MWR-D-15-0242.1.

Berner, J., S.-Y. Ha, J. P. Hacker, A. Fournier, and C. Snyder, 2011: Model uncertainty in a mesoscale ensemble prediction system: Stochastic versus multiphysics representations. *Mon. Wea. Rev.*, **139**, 1972–1995, doi:10.1175/2010MWR3595.1.

——, K. R. Fossell, S.-Y. Ha, J. P. Hacker, and C. Snyder, 2015: Increasing the skill of probabilistic forecasts: Understanding performance improvements from model-error representations. *Mon. Wea. Rev.*, **143**, 1295–1320, doi:10.1175/MWR-D-14-00091.1.

Bishop, C. H., B. J. Etherton, and S. J. Majumdar, 2001: Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. *Mon. Wea. Rev.*, **129**, 420–436, doi:10.1175/1520-0493(2001)129<0420:ASWTET>2.0.CO;2.

Brennan, M. J., and G. M. Lackmann, 2005: The influence of incipient latent heat release on the precipitation distribution of the 24–25 January 2000 U.S. East Coast cyclone. *Mon. Wea. Rev.*, **133**, 1913–1937, doi:10.1175/MWR2959.1.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Brown, L. D., T. T. Cai, and A. DasGupta, 2001: Interval estimation for a binomial proportion. *Stat. Sci.*, **16**, 101–133.

Carr, F., and R. Rood, 2015: Report of the UCACN Model Advisory Committee. UCAR Community Advisory Committee for NCEP, 72 pp. [Available online at http://www.ncep.noaa.gov/director/ucar_reports/ucacn_20151207/UMAC_Final_Report_20151207-v14.pdf.]

Charles, M. E., and B. A. Colle, 2009a: Verification of extratropical cyclones within the NCEP operational models. Part I: Analysis errors and short-term NAM and GFS forecasts. *Wea. Forecasting*, **24**, 1173–1190, doi:10.1175/2009WAF2222169.1.

——, and ——, 2009b: Verification of extratropical cyclones within the NCEP operational models. Part II: The short-range ensemble forecast system. *Wea. Forecasting*, **24**, 1191–1214, doi:10.1175/2009WAF2222170.1.

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, doi:10.1175/1520-0493(2001)129<0569:CAALSH>2.0.CO;2.

DeGaetano, A. T., M. E. Hirsch, and S. J. Colucci, 2002: Statistical prediction of seasonal East Coast winter storm frequency. *J. Climate*, **15**, 1101–1117, doi:10.1175/1520-0442(2002)015<1101:SPOSEC>2.0.CO;2.

Doesken, N. J., and D. A. Robinson, 2009: The challenge of snow measurements. *Historical Climate Variability and Impacts in North America*, L.-A. Dupigny-Giroux and C. J. Mock, Eds., Springer, 251–273, doi:10.1007/978-90-481-2828-0_15.

Du, J., G. DiMego, B. Zhou, D. Jovic, B. Ferrier, and B. Yang, 2015: Short Range Ensemble Forecast (SREF) system at NCEP: Recent development and future transition. *23rd Conf. on Numerical Weather Prediction/27th Conf. on Weather Analysis and Forecasting*, Chicago, IL, Amer. Meteor. Soc., 2A.5. [Available online at https://ams.confex.com/ams/27WAF23NWP/webprogram/Paper273421.html.]

Dudhia, J., 1989: Numerical study of convection observed during the Winter Monsoon Experiment using a mesoscale two-dimensional model. *J. Atmos. Sci.*, **46**, 3077–3107, doi:10.1175/1520-0469(1989)046<3077:NSOCOD>2.0.CO;2.

Evans, M., and M. L. Jurewicz, 2009: Correlations between analyses and forecasts of banded heavy snow ingredients and observed snowfall. *Wea. Forecasting*, **24**, 337–350, doi:10.1175/2008WAF2007105.1.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.*, **99**, 10 143–10 162, doi:10.1029/94JC00572.

Ganetis, S. A., and B. A. Colle, 2015: The thermodynamic and microphysical evolution of an intense snowband during the northeast U.S. blizzard of 8–9 February 2013. *Mon. Wea. Rev.*, **143**, 4104–4125, doi:10.1175/MWR-D-14-00407.1.

Gaza, B., and L. F. Bosart, 1990: Trough merger characteristics over North America. *Wea. Forecasting*, **5**, 314–331, doi:10.1175/1520-0434(1990)005<0314:TMCONA>2.0.CO;2.

Grell, G. A., and D. Dévényi, 2002: A generalized approach to parameterizing convection combining ensemble and data assimilation techniques. *Geophys. Res. Lett.*, **20**, 1693, doi:10.1029/2002GL015311.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559–570, doi:10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Hirsch, M. E., A. T. DeGaetano, and S. J. Colucci, 2001: An East Coast winter storm climatology. *J. Climate*, **14**, 882–899, doi:10.1175/1520-0442(2001)014<0882:AECWSC>2.0.CO;2.

Hutchinson, T. A., 1995: An analysis of the NMC's nested grid model forecasts of Alberta Clippers. *Wea. Forecasting*, **10**, 632–641, doi:10.1175/1520-0434(1995)010<0632:AAONNG>2.0.CO;2.

Janjić, Z. I., 1994: The step-mountain eta coordinate model: Further developments of the convection, viscous sublayer, and turbulence closure schemes. *Mon. Wea. Rev.*, **122**, 927–945, doi:10.1175/1520-0493(1994)122<0927:TSMECM>2.0.CO;2.

——, 1996: The surface layer in the NCEP Eta Model. Preprints, *11th Conf. on Numerical Weather Prediction*, Norfolk, VA, Amer. Meteor. Soc., 354–355.

——, 2002: Nonsingular implementation of the Mellor–Yamada level 2.5 scheme in the NCEP Meso model. NCEP Office Note 437, National Centers for Environmental Prediction, 61 pp. [Available online at http://www2.mmm.ucar.edu/wrf/users/phys_refs/SURFACE_LAYER/eta_part4.pdf.]

Kocin, P. J., and L. W. Uccellini, 2004a: A snowfall impact scale derived from Northeast storm snowfall distributions. *Bull. Amer. Meteor. Soc.*, **85**, 177–194, doi:10.1175/BAMS-85-2-177.

——, and ——, 2004b: *Northeast Snowstorms*. Vol. 1. *Meteor. Monogr.*, No. 54, Amer. Meteor. Soc., 296 pp.

Kumjian, M. R., and K. A. Lombardo, 2017: Insights into the evolving microphysical and kinematic structure of northeastern U.S. winter storms from dual-polarization Doppler radar. *Mon. Wea. Rev.*, **145**, 1033–1061, doi:10.1175/MWR-D-15-0451.1.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.*, **20**, 130–131, doi:10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

Maltzahn, C., and Coauthors, 2016: Big Weather Web: A common and sustainable big data infrastructure in support of weather prediction research and education in universities. [Available online at http://bigweatherweb.org.]

Mass, C. F., D. Owens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? The results of two years of real-time numerical weather prediction over the Pacific Northwest. *Bull. Amer. Meteor. Soc.*, **83**, 407–430, doi:10.1175/1520-0477(2002)083<0407:DIHRPM>2.3.CO;2.

Matheson, J. E., and R. L. Winkler, 1976: Scoring rules for continuous probability distributions. *Manage. Sci.*, **22**, 1087–1095, doi:10.1287/mnsc.22.10.1087.

Melhauser, C., and F. Zhang, 2012: Practical and intrinsic predictability of severe and convective weather at the mesoscales. *J. Atmos. Sci.*, **69**, 3350–3371, doi:10.1175/JAS-D-11-0315.1.

Miller, J. E., 1946: Cyclogenesis in the Atlantic coastal region of the United States. *J. Meteor.*, **3**, 31–44, doi:10.1175/1520-0469(1946)003<0031:CITACR>2.0.CO;2.

Mlawer, E. J., S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough, 1997: Radiative transfer for inhomogeneous atmosphere: RRTM, a validated correlated-k model for the long wave. *J. Geophys. Res.*, **102**, 16 663–16 682, doi:10.1029/97JD00237.

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 intercomparison. *Wea. Forecasting*, **31**, 371–394, doi:10.1175/WAF-D-14-00112.1.

Nicosia, D. J., and R. H. Grumm, 1999: Mesoscale band formation in three major northeastern United States snowstorms. *Wea. Forecasting*, **14**, 346–368, doi:10.1175/1520-0434(1999)014<0346:MBFITM>2.0.CO;2.

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

——, J. S. Waldstreicher, D. Keyser, and L. F. Bosart, 2006: A forecast strategy for anticipating cold season mesoscale band formation within eastern U.S. cyclones. *Wea. Forecasting*, **21**, 3–23, doi:10.1175/WAF907.1.

——, B. A. Colle, and S. E. Yuter, 2008: High-resolution observations and model simulations of the life cycle of an intense mesoscale snowband over the northeastern United States. *Mon. Wea. Rev.*, **136**, 1433–1456, doi:10.1175/2007MWR2233.1.

——, K. F. Brill, and W. A. Hogsett, 2014: Using percentiles to communicate snowfall uncertainty. *Wea. Forecasting*, **29**, 1259–1265, doi:10.1175/WAF-D-14-00019.1.

Ota, Y., J. C. Derber, E. Kalnay, and T. Miyoshi, 2013: Ensemble-based observation impact estimates using the NCEP GFS. *Tellus*, **65A**, 20038, doi:10.3402/tellusa.v65i0.20038.

Palmer, T. N., R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G. J. Shutts, M. Steinheimer, and A. Weisheimer, 2009: Stochastic parametrization and model uncertainty. ECMWF Tech. Memo. 598, 42 pp. [Available online at http://www.ecmwf.int/sites/default/files/elibrary/2009/11577-stochastic-parametrization-and-model-uncertainty.pdf.]

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, doi:10.1175/MWR2906.1.

Roebber, P., S. Bruening, D. Schultz, and J. Cortinas, 2003: Improving snowfall forecasting by diagnosing snow density. *Wea. Forecasting*, **18**, 264–287, doi:10.1175/1520-0434(2003)018<0264:ISFBDS>2.0.CO;2.

Romine, G. S., C. S. Schwartz, J. Berner, K. R. Fossell, C. Snyder, J. L. Anderson, and M. L. Weisman, 2014: Representing forecast error in a convection-permitting ensemble system. *Mon. Wea. Rev.*, **142**, 4519–4541, doi:10.1175/MWR-D-14-00100.1.

Root, B., P. Knight, G. Young, S. Greybush, R. Grumm, R. Holmes, and J. Ross, 2007: A fingerprinting technique for major weather events. *J. Appl. Meteor. Climatol.*, **46**, 1053–1066, doi:10.1175/JAM2509.1.

Sanders, F., and L. F. Bosart, 1985: Mesoscale structure in the megalopolitan snowstorm of 11–12 February 1983. Part I: Frontogenetical forcing and symmetric instability. *J. Atmos. Sci.*, **42**, 1050–1061, doi:10.1175/1520-0469(1985)042<1050:MSITMS>2.0.CO;2.

Scheuerer, M., and T. M. Hamill, 2015: Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Mon. Wea. Rev.*, **143**, 4578–4596, doi:10.1175/MWR-D-15-0061.1.

Schwartz, C. S., G. S. Romine, R. A. Sobash, K. R. Fossell, and M. L. Weisman, 2015: NCAR's experimental real-time convection-allowing ensemble prediction system. *Wea. Forecasting*, **30**, 1645–1654, doi:10.1175/WAF-D-15-0103.1.

Seo, D. J., 1998: Real-time estimation of rainfall fields using rain gauge data under fractional coverage conditions. *J. Hydrol.*, **208**, 25–36, doi:10.1016/S0022-1694(98)00140-1.

Shutts, G. J., 2005: A kinetic energy backscatter algorithm for use in ensemble prediction systems. *Quart. J. Roy. Meteor. Soc.*, **131**, 3079–3102, doi:10.1256/qj.04.106.

Siddique, R., A. Meija, J. Brown, S. Reed, and P. Ahnert, 2015: Verification of precipitation forecasts from two numerical weather prediction models in the middle Atlantic region of the USA: A precursory analysis to hydrologic forecasting. *J. Hydrol.*, **529**, 1390–1406, doi:10.1016/j.jhydrol.2015.08.042.

Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-475+STR, 113 pp., doi:10.5065/D68S4MVH.

Stensrud, D. J., J.-W. Bao, and T. T. Warner, 2000: Using initial condition and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. *Mon. Wea. Rev.*, **128**, 2077–2107, doi:10.1175/1520-0493(2000)128<2077:UICAMP>2.0.CO;2.

——, and Coauthors, 2009: Convective-scale Warn-on-Forecast system. *Bull. Amer. Meteor. Soc.*, **90**, 1487–1499, doi:10.1175/2009BAMS2795.1.

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, doi:10.1175/2008MWR2387.1.

Torn, R. D., and G. J. Hakim, 2008: Ensemble-based sensitivity analysis. *Mon. Wea. Rev.*, **136**, 663–677, doi:10.1175/2007MWR2132.1.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317–2330, doi:10.1175/1520-0477(1993)074<2317:EFANTG>2.0.CO;2.

Tracton, S. M., and E. Kalnay, 1993: Operational ensemble prediction at the National Meteorological Center: Practical aspects. *Wea. Forecasting*, **8**, 379–398, doi:10.1175/1520-0434(1993)008<0379:OEPATN>2.0.CO;2.

Wang, X., D. Parrish, D. Kleist, and J. Whitaker, 2013: GSI 3DVar-based ensemble–variational hybrid data assimilation for NCEP Global Forecast System: Single-resolution experiments. *Mon. Wea. Rev.*, **141**, 4098–4117, doi:10.1175/MWR-D-12-00141.1.

Wei, M., Z. Toth, R. Wobus, and Y. Zhu, 2008: Initial perturbations based on the ensemble transform (ET) technique in the NCEP global operational forecast system. *Tellus*, **60A**, 62–79, doi:10.3402/tellusa.v65i0.20038.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. Elsevier, 676 pp.

Zhang, F., C. Snyder, and R. Rotunno, 2002: Mesoscale predictability of the "surprise" snowstorm of 24–25 January 2000. *Mon. Wea. Rev.*, **130**, 1617–1632, doi:10.1175/1520-0493(2002)130<1617:MPOTSS>2.0.CO;2.

——, ——, and ——, 2003: Effects of moist convection on mesoscale predictability. *J. Atmos. Sci.*, **60**, 1173–1184, doi:10.1175/1520-0469(2003)060<1173:EOMCOM>2.0.CO;2.

——, N. Bei, R. Rotunno, C. Snyder, and C. C. Epifanio, 2007: Mesoscale predictability of moist baroclinic waves: Cloud-resolving experiments and multistage error growth dynamics. *J. Atmos. Sci.*, **64**, 3579–3594, doi:10.1175/JAS4028.1.

——, F. Weng, J. A. Sippel, Z. Meng, and C. H. Bishop, 2009: Cloud-resolving hurricane initialization and prediction through assimilation of Doppler radar observations with an ensemble Kalman filter. *Mon. Wea. Rev.*, **137**, 2105–2125, doi:10.1175/2009MWR2645.1.

Zheng, M., E. K. M. Chang, and B. A. Colle, 2013: Ensemble sensitivity tools for assessing extratropical cyclone intensity and track predictability. *Wea. Forecasting*, **28**, 1133–1156, doi:10.1175/WAF-D-12-00132.1.