The Impacts of Using Mixed Physics in the Community Leveraged Unified Ensemble

William A. Gallus Jr.

Department of Geological and Atmospheric Sciences, Iowa State University, Ames, Iowa

Jamie Wolff, John Halley Gotway, Michelle Harrold, and Lindsay Blank

National Center for Atmospheric Research/Developmental Testbed Center/Joint Numerical Testbed, Boulder, Colorado

Jeff Beck

Cooperative Institute for Research in the Atmosphere, NOAA/OAR/ESRL/GSD/MDB, Boulder, Colorado

(Manuscript received 4 December 2018, in final form 30 March 2019)

ABSTRACT

A well-known problem in high-resolution ensembles has been a lack of sufficient spread among members. Modelers often have used mixed physics to increase spread, but this can introduce problems including computational expense, clustering of members, and members that are not all equally skillful. Thus, a detailed examination of the impacts of using mixed physics is important. The present study uses two years of Community Leveraged Unified Ensemble (CLUE) output to isolate the impact of mixed physics in 36-h forecasts made using a convection-permitting ensemble with 3-km horizontal grid spacing. One 10-member subset of the CLUE used only perturbed initial conditions (ICs) and lateral boundary conditions (LBCs) while another 10-member ensemble used the same mixed ICs and LBCs but also introduced mixed physics. The cases examined occurred during NOAA's Hazardous Weather Testbed Spring Forecast Experiments in 2016 and 2017. Traditional gridpoint metrics applied to each member and the ensemble as a whole, along with object-based verification statistics for all members, were computed for composite reflectivity and 1- and 3-h accumulated precipitation using the Model Evaluation Tools (MET) software package. It is found that the mixed physics increases variability substantially among the ensemble members, more so for reflectivity than precipitation, such that the envelope of members is more likely to encompass the observations. However, the increased variability is mostly due to the introduction of both substantial high biases in members using one microphysical scheme, and low biases in other schemes. Overall ensemble skill is not substantially different from the ensemble using a single physics package.

1. Introduction

Because of the uncertainty present in weather forecasts, ensemble forecasting has become an essential part of operational forecasting (e.g., Tracton and Kalnay 1993; Molteni et al. 1996; Du et al. 2003; Buizza et al. 2007). At first, ensemble systems were introduced to provide additional guidance, such as a measure of uncertainty and mean values, for global-scale forecasts (e.g., Toth and Kalnay 1993) emphasizing medium and longer time ranges, with the ensemble members using perturbed initial conditions (ICs), and when applied to regional domains, perturbed lateral boundary conditions (LBCs). Increasingly often, ensemble systems are being used with convection-allowing horizontal grid spacing with applications toward quantitative precipitation forecasts (QPF) and severe weather (e.g., Clark et al. 2012; Gallo et al. 2016; Clark et al. 2018).

It has been shown that some of the same techniques used to create the ensemble members for global models through perturbed IC/LBCs do not provide enough spread for the high-resolution forecasts that emphasize shorter time ranges, and thus these ensemble systems have often used mixed physics in an effort to increase the spread in these forecasts (e.g., Stensrud et al. 2000; Hacker et al. 2011; Berner et al. 2011, 2015). Although the increase in spread may result in a more useful ensemble forecast better able to capture the observed precipitation or severe weather-producing event, there

Corresponding author: William A. Gallus Jr., wgallus@iastate.edu

DOI: 10.1175/WAF-D-18-0197.1

TABLE 1. Specifications for the S-Phys single physics ensemble in 2016. NAM refers to 12-km NAM output with ''a'' being analysis and ''f'' forecast; 3DVAR refers to ARPS3DVAR and cloud analysis. Model names appended with ''pert'' refer to perturbations extracted from a 16-km grid-spacing SREF member.

Member	IC	LBC	Microphysics	LSM	PBL	Model
1	NAMa+3DVAR	NAMf	Thompson	Noah	MYJ	arw
2	1 + arw-p1_pert	arw-p1	Thompson	Noah	MYJ	arw
3	1 + arw-n1_pert	arw-n1	Thompson	Noah	MYJ	arw
4	1 + arw-p2_pert	arw-p2	Thompson	Noah	MYJ	arw
5	1 + arw-n2_pert	arw-n2	Thompson	Noah	MYJ	arw
7	1 + nnmb-p1_pert	nmmb-p1	Thompson	Noah	MYJ	arw
8	1 + nmmb-n1_pert	nmmb-n1	Thompson	Noah	MYJ	arw
9	1 + nmmb-p2_pert	nmmb-p2	Thompson	Noah	MYJ	arw
10	1 + nmmb-n2_pert	nmmb-n2	Thompson	Noah	MYJ	arw

are theoretical and practical disadvantages to using this mixed physics approach. Often, the different physics schemes result in systematic biases (e.g., Jankov et al. 2017), and clustering of members can occur (Johnson et al. 2011) where members using, for instance, the same microphysics schemes resemble each other more than they resemble the observations or any other members using different microphysics schemes (e.g., Stensrud et al. 2000). In addition, development and maintenance of a suite of different physics schemes is resource intensive.

Because techniques such as ensemble Kalman filters (e.g., Houtekamer and Mitchell 1998; Johnson et al. 2015) have gained use in recent years to create perturbed IC/LBCs, it is worth exploring whether the use of mixed physics packages in convection-allowing ensembles provides enough benefits to justify the continued use of this approach in spite of the problems. The Community Leveraged Unified Ensemble (CLUE), first employed to assist the NOAA Hazardous Weather Testbed Spring Forecast Experiment (HWT-SFE) in 2016, is a collaboratively run ensemble of over 60 members, designed to allow exploration of questions relating to ensemble construction (Clark et al. 2018) such as ''what is the impact of adding mixed physics to an ensemble already making use of perturbed IC/LBCs?'' The present study uses 3-km CLUE output from both 2016 and 2017 from the two subensembles that both use perturbed IC/LBCs with one also including mixed physics to determine the impact of using mixed physics in a convection-allowing grid spacing ensemble. Section 2 discusses the methodology, section 3 presents the results, and section 4 offers the conclusions and discussion.

2. Methodology

To explore the impact of mixed physics within an ensemble, 9 members of a CLUE subensemble using the same physics schemes but with member variability coming from IC and LBC perturbations (hereafter known as S-Phys, see Table 1) and 9 members of a different CLUE subensemble using the same IC/LBC perturbations as S-Phys but also employing mixed physics (hereafter known as Core, see Table 2) were examined from the NOAA HWT-SFE in 2016. Additionally, 10 members from similar ensembles were compared in 2017 (Tables 3 and 4). Although the ensembles were designed to contain 10 members in both years, in 2016 S-Phys was missing member 6, thus, member 2 was eliminated from Core to allow an equal number of members to be compared. This particular Core member was chosen to be neglected because it was the only member not using the North American Mesoscale Model (NAM) for its ICs and LBCs and the Noah

TABLE 2. Specifications for the Core mixed physics ensemble in 2016. Notations as in Table 1.

Member	IC	LBC	Microphysics	LSM	PBL	Model
1	NAMa+3DVAR	NAMf	Thompson	Noah	MYJ	arw
3	1 + arw-p1_pert	arw-p1	P3	Noah	YSU	arw
4	1 + arw-n1_pert	arw-n1	MY	Noah	MYNN	arw
5	1 + arw-p2_pert	arw-p2	Morrison	Noah	MYJ	arw
6	1 + arw-n2_pert	arw-n2	P3	Noah	YSU	arw
7	1 + nnmb-p1_pert	nmmb-p1	MY	Noah	MYNN	arw
8	1 + nmmb-n1_pert	nmmb-n1	Morrison	Noah	YSU	arw
9	1 + nmmb-p2_pert	nmmb-p2	P3	Noah	MYJ	arw
10	1 + nmmb-n2_pert	nmmb-n2	Thompson	Noah	MYNN	arw

TABLE 3. Specifications for the S-Phys single physics ensemble in 2017. Notation same as in Table 1, with RAPa referring to 13-km RAP analysis, and GFSf referring to 1800 UTC initialized GFS forecasts.

Member	IC	LBC	Microphysics	LSM	PBL	Model
1	RAPa+3DVAR	GFSf	Thompson	RUC	MYNN	arw
2	NAMa+3DVAR	NAMf	Thompson	RUC	MYNN	arw
3	1 + arw-p1_pert	arw-p1	Thompson	RUC	MYNN	arw
4	1 + arw-n1_pert	arw-n1	Thompson	RUC	MYNN	arw
5	1 + nmmb-p1_pert	nmmb-p1	Thompson	RUC	MYNN	arw
6	1 + nmmb-n1_pert	nmmb-n1	Thompson	RUC	MYNN	arw
7	2 + arw-p2_pert	arw-p2	Thompson	RUC	MYNN	arw
8	2 + arw-n2_pert	arw-n2	Thompson	RUC	MYNN	arw
9	2 + nmmb-p2_pert	nmmb-p2	Thompson	RUC	MYNN	arw
10	2 + nmmb-n2_pert	nmmb-n2	Thompson	RUC	MYNN	arw

land surface model (Mitchell et al. 2001). All CLUE ensemble members used the Weather Research and Forecasting (WRF) Model (Skamarock et al. 2008) with the Advanced Research version of WRF (ARW) dynamic core over a continental United States domain having 3-km horizontal grid spacing. No convective parameterization was used. Simulations were initialized at 0000 UTC for all cases and integrated for 36 h. In 2016, the S-Phys ensemble used the Thompson (Thompson et al. 2008) microphysics with the Noah LSM and Mellor–Yamada–Janjić (MYJ; Janjić 1994) PBL schemes. These were also the schemes used in the control member within Core. In Core, the varied microphysics included the Predicted Particle Property (P3; Morrison and Milbrandt 2015), Milbrandt–Yau (MY; Milbrandt et al. 2008) and Morrison (Morrison et al. 2009) schemes, and PBL scheme variations included Mellor–Yamada–Nikanishi–Niino (MYNN; Nakanishi and Niino 2009) and Yonsei University (YSU; Hong et al. 2006). Both ensembles used a mixture of initial conditions and lateral boundary conditions. In 2016, all members used in the present study were initialized using the NAM and radar data assimilation via the ARPS 3DVAR (Xue et al. 2003, Hu et al. 2006) system, but with variations from the control member coming through use of perturbations from the Short-Range Ensemble Forecast (SREF; Du et al. 2003) added to the control initial conditions. Similar IC/LBCs were used for members 1 and 3–6 in 2017. Conversely, Core members 2 and 7–10 in 2017 differed in their IC/LBCs by using Rapid Refresh (RAP) (Benjamin et al. 2016) analyses with the Global Forecasting System (GFS) supplying the LBCs. Also noteworthy for 2017 was the switch to the MYNN PBL scheme and RUC land surface scheme (Smirnova et al. 1997) in all S-Phys members.

To evaluate the impacts of using mixed physics, Model Evaluation Tools (MET) software package version 6.1 (Bullock et al. 2017) and METviewer (a database and display system) were used to evaluate the ensembles. Verification was performed using Multi-Radar Multi-Sensor (MRMS; Zhang et al. 2016) observations. MRMS uses radar-based data integrated with surface and upper-air observations, satellite data, lightning observations, and rain gauge observations to generate a suite of severe weather and quantitative precipitation estimation (QPE) products at very high spatial (1 km) resolution (Zhang et al. 2016). The MRMS data were regridded to the model integration domain to allow for grid-to-grid comparisons. Budget interpolation was used for the QPE field with nearest neighbor employed for the composite reflectivity regridding. The budget

TABLE 4. Specifications for the Core mixed physics ensemble in 2017. Notations as in Table 3.

Member	IC	LBC	Microphysics	LSM	PBL	Model
1	NAMa+3DVAR	NAMf	Thompson	Noah	MYJ	arw
2	RAPa+3DVAR	GFSf	Thompson	RUC	MYNN	arw
3	1 + arw-p1_pert	arw-p1	P3	Noah	YSU	arw
4	1 + arw-n1_pert	arw-n1	MY	Noah	MYNN	arw
5	1 + nmmb-p1_pert	nmmb-p1	Morrison	Noah	MYJ	arw
6	1 + nmmb-n1_pert	nmmb-n1	P3	Noah	YSU	arw
7	2 + arw-p2_pert	arw-p2	MY	Noah	MYNN	arw
8	2 + arw-n2_pert	arw-n2	Morrison	Noah	YSU	arw
9	2 + nmmb-p2_pert	nmmb-p2	P3	Noah	MYJ	arw
10	2 + nmmb-n2_pert	nmmb-n2	Thompson	Noah	MYNN	arw

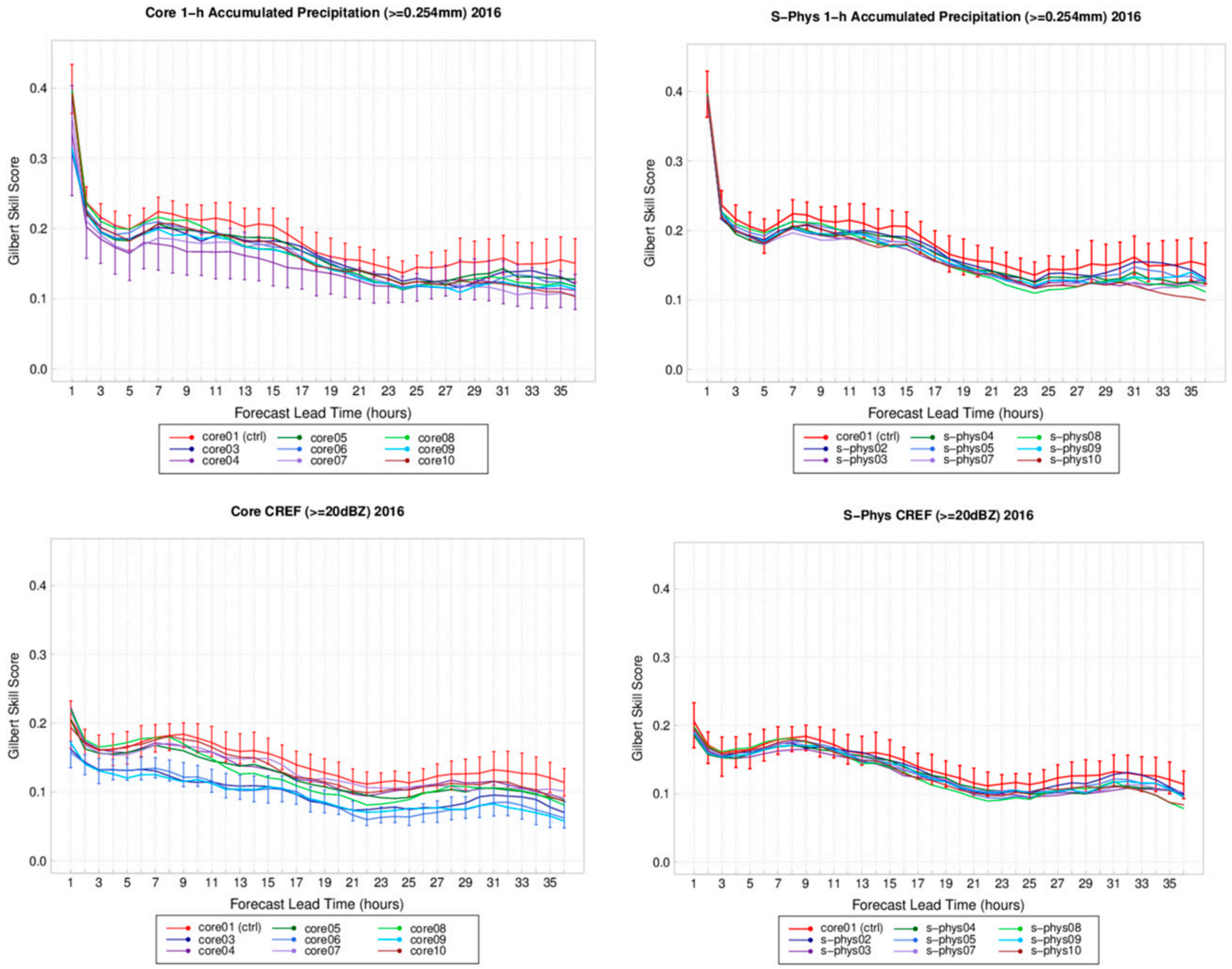


FIG. 1. GSS for each member of (left) Core and (right) S-Phys for 2016 for (top) 1-h precipitation ≥ 0.254 mm and (bottom) CREF ≥ 20 dBZ. In Core, red indicates member 1, dark blue 3, dark purple 4, dark green 5, blue 6, light purple 7, light green 8, light blue 9, and dark red 10 (see Table 2 for configuration details). In S-Phys, red is member 1, dark blue 2, dark purple 3, dark green 4, blue 5, light purple 7, light green 8, light blue 9, and dark red 10 (see Table 1 for configuration details). Colors for Core are grouped by microphysics scheme: Thompson in shades of red, P3 in shades of blue, MY in shades of purple, and Morrison in shades of green. The vertical bars represent 95% CIs for selected curves (core01 in all, core04 in top left, and core06 in bottom left).

interpolation method, also known as nearest-neighbor averaging, is a way of conserving the total area-average variable value (Wolff et al. 2014). In the present study, three fields were assessed, including 1-h precipitation, 3-h precipitation, and composite reflectivity (CREF) computed directly in WRF so as to be consistent with the assumptions used in each microphysical scheme.

Several types of verification metrics were applied. A deterministic verification using traditional grid-to-grid comparisons, including Gilbert skill score (GSS; Schaefer 1990) and frequency bias, was computed for each member of the two ensembles. Averages were taken of the members to evaluate how the mixed physics might be impacting general skill and areal coverage of reflectivity or precipitation above specified thresholds within its members compared to the S-Phys members.

GSS is the fraction of observed and/or forecast events that were correctly predicted, adjusted for hits associated with random chance. GSS values can range from −1/3 to 1, with a no-skill forecast having a value of 0 and a perfect forecast being equal to 1. Frequency bias is the ratio of the frequency of forecast events to observed events (or total forecast area divided by total observed area) and indicates whether there is an underforecast (<1) or overforecast (>1) of an event; an unbiased forecast has a frequency bias of 1.

In addition, object-based spatial verification was performed on each member using the Method for Object-based Diagnostic Evaluation (MODE; Davis et al. 2006a, 2006b, 2009) tool, and averages of MODE

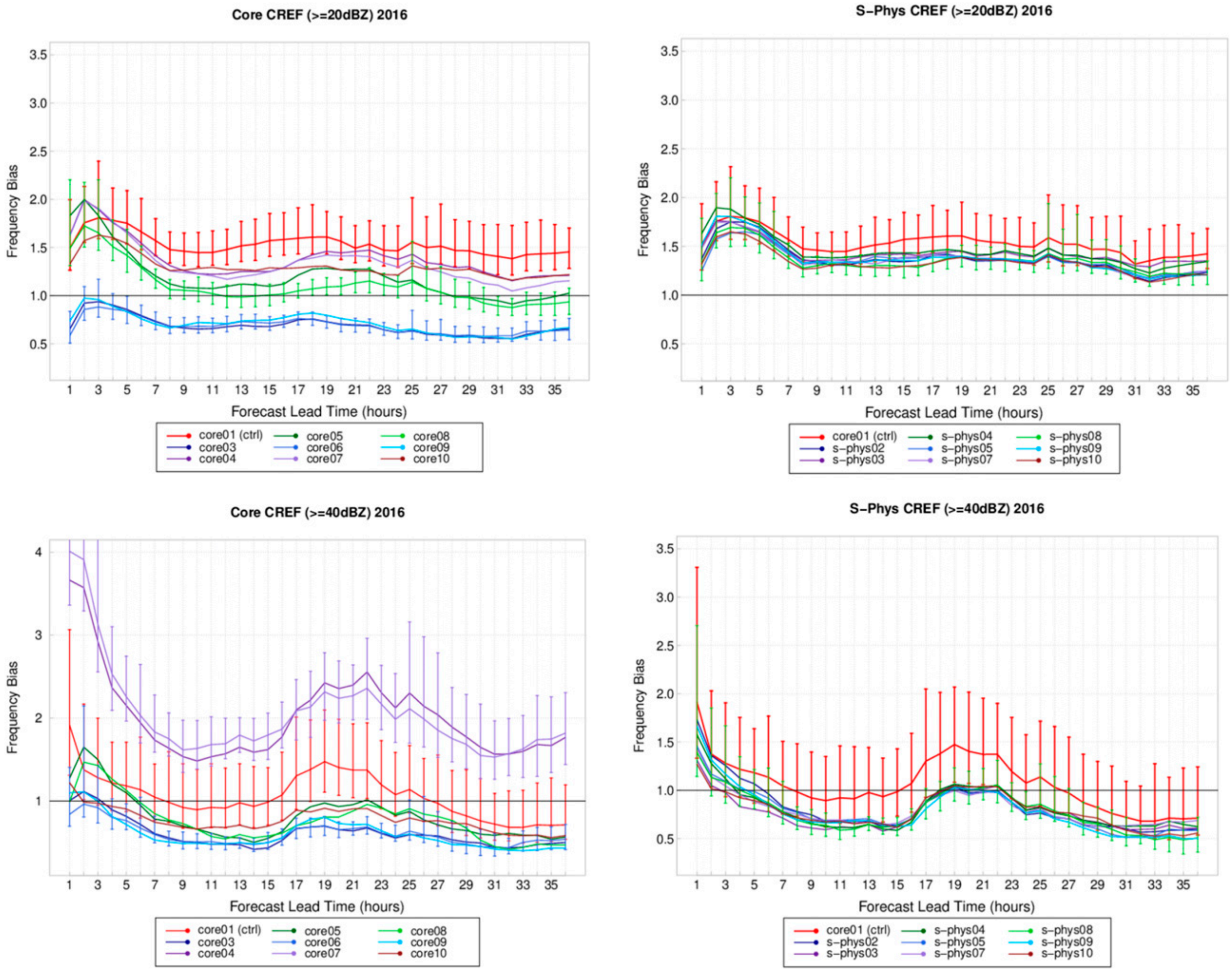


FIG. 2. Frequency bias for 2016 output for (left) Core and (right) S-Phys for (top) ≥20- and (bottom) ≥40-dBZ CREF threshold. Individual Core member and S-Phys member physics schemes are as indicated in Fig. 1. The vertical bars represent 95% CIs for selected curves (core01 in all, core06 and core08 in top left, s-phys08 in right panels, and core06 and core07 in bottom right).

attributes were computed for each ensemble, loosely following Gallus (2010). These attributes included the area and displacement of the objects, median and 90th percentile values, intensity sum, and counts of objects. Objects were defined using 1- and 3-h accumulated rainfall thresholds of 0.254, 2.54, and 6.35 mm at grid points, and composite reflectivity thresholds of 20, 30, and 40 dBZ. Finally, standard ensemble verification was performed on the probabilistic forecasts using measures such as receiver operating characteristic (ROC) areas, reliability (measure of the average difference between forecast probability and observed frequency), and Brier score (measure of the mean squared probability error).

For most of the metrics described above, confidence intervals (CIs) at the 95% level were applied to estimate the uncertainty associated with sampling variability. A conservative estimate of statistical significance can then be used whereby differences are statistically significant at the 95% level if the confidence intervals associated with different ensembles or individual members do not overlap. This method was used for frequency bias and all MODE attributes. However, for GSS, a more robust pairwise difference approach was applied to measure statistical significance. The CIs were computed using the appropriate statistical method (Gilleland 2010), with bootstrapping used for GSS and frequency bias, and standard error about the median for all MODE attributes. For the standard error algorithm, a normal distribution is assumed and the variance of the sample is considered. Bootstrapping provides an estimate of uncertainty using a numerical resampling method. In the present study, resampling with replacement was performed 1000 times. Observational uncertainty was not considered in this study.

The HWT-SFE 2016 ran from 2 May to 3 June, with model output only available from the weekday portions

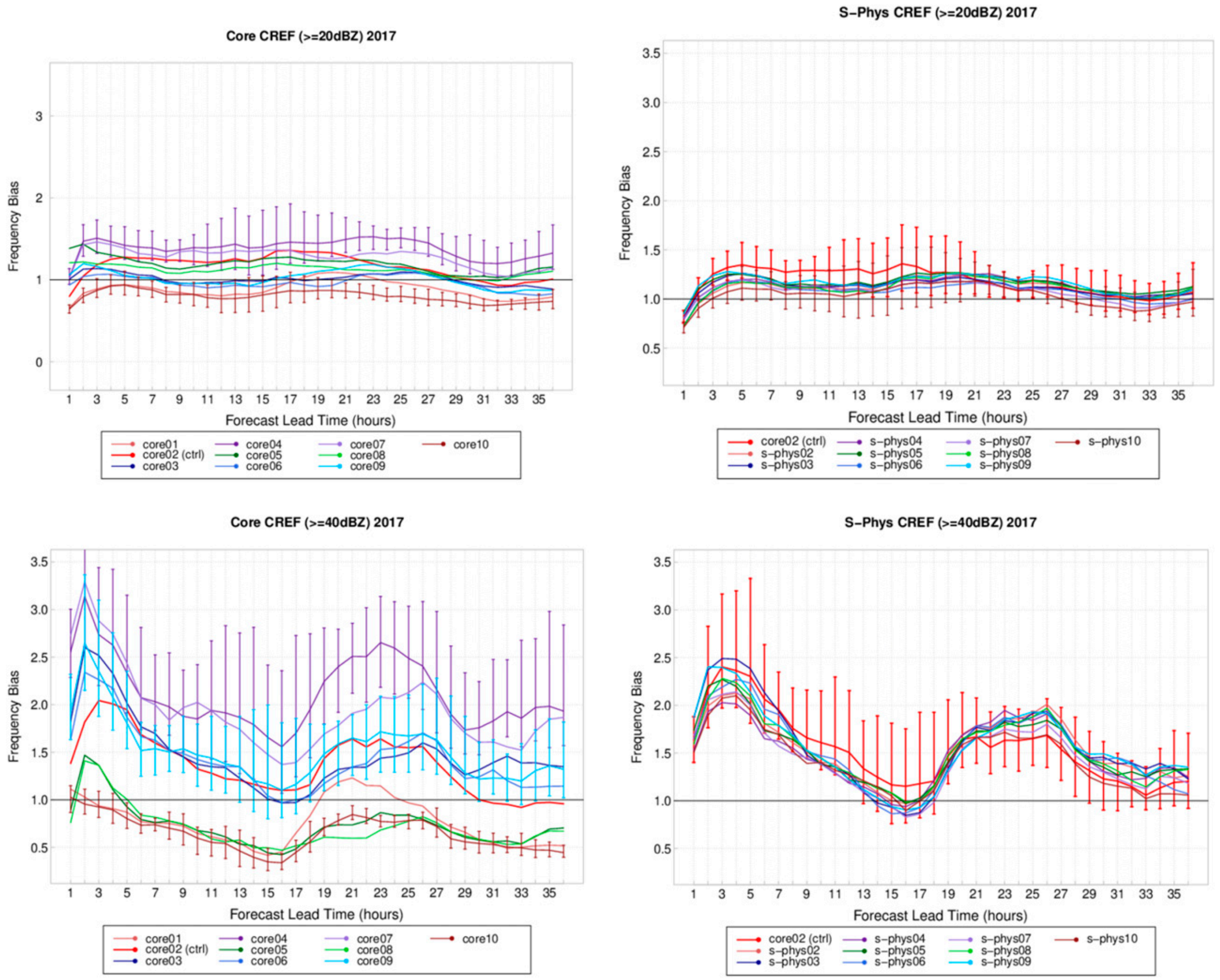


FIG. 3. As in Fig. 2, but for 2017. In Core, light red indicates member 1, red 2, dark blue 3, dark purple 4, dark green 5, blue 6, light purple 7, light green 8, light blue 9, and dark red 10 (see Table 4 for configuration details). In S-Phys, red is Core member 2, light red 2, dark blue 3, dark purple 4, dark green 5, blue 6, light purple 7, light green 8, light blue 9, and dark red 10 (see Table 3 for configuration details). Colors for Core are grouped by microphysics scheme: Thompson in shades of red, P3 in shades of blue, MY in shades of purple, and Morrison in shades of green. The vertical bars represent 95% CIs for selected curves (core07 and core10 in top left, core02 and s-phys10 in top right, core07, core09, and core10 in bottom left, and core02 in bottom right).

of that period. Similarly, in 2017 the project ran from 1 May to 2 June during the weekdays. Because some of the composite reflectivity data from runs using the MY microphysics scheme in 2016 was corrupted, the size of the dataset that could be used for CREF was reduced. Likewise, in 2017, a problem prevented S-Phys from being run during the first part of the project. In the end, a total of 22 cases were available for comparison of precipitation data, 17 cases for comparison of CREF in 2016, and only 12 cases in 2017 for both fields (the case size represents events for which output was available from both ensembles in each year).

In addition to the comparisons that could be performed using MET, one other comparison was made for the 2016 ensembles. For a subset of 10 cases with relatively pristine convective initiation (new convection forming at least 100 km away from existing convection), differences in the ensemble prediction of this initiation were evaluated. Location and timing were studied using each member of both ensembles to understand differences in both skill and variability of solutions.

3. Results

The impacts of using mixed physics were determined by applying multiple verification strategies including grid-to-grid measures applied to individual members, the same metrics averaged for all ensemble members,

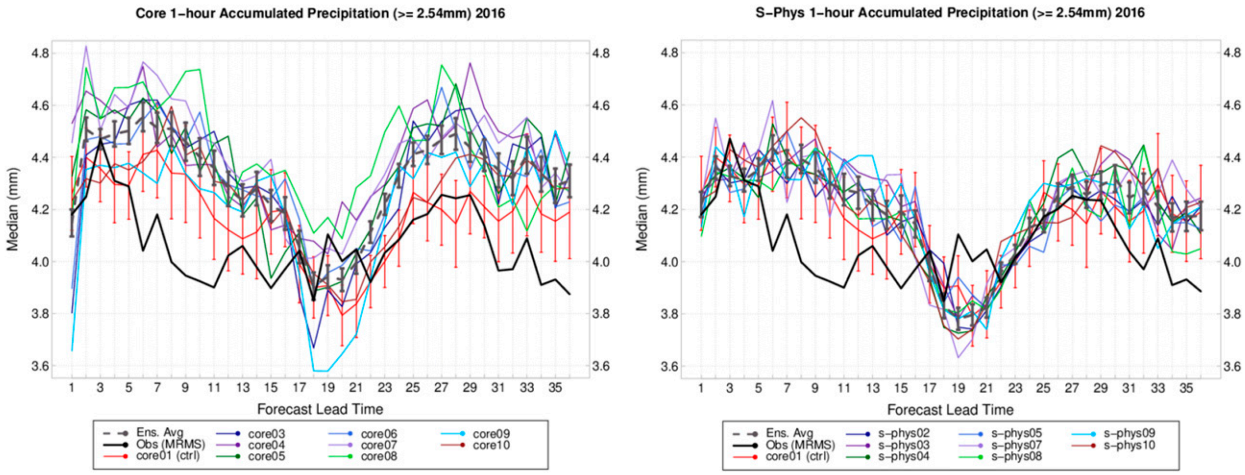


FIG. 4. Median 1-h precipitation value (mm) for both ensembles in 2016 for (left) Core and (right) S-Phys for the ≥2.54-mm threshold. Colors of curves representing different members follow same notation as in Fig. 1, with observations in black and the ensemble average represented by a dashed dark gray line. The vertical bars represent 95% CIs for selected curves (core01 and ensemble average in both).

MODE attributes, and traditional ensemble metrics making use of probability values. In the discussion below, emphasis will be on 1-h precipitation and CREF. In general, 3-h precipitation behaved similarly to 1-h precipitation, except results exhibited more skill as would be expected for a longer time period.

a. *Grid-to-grid metrics*

GSS computed for both a threshold of ≥0.254 mm for 1-h precipitation and ≥20 dBZ for CREF for the 2016 output is shown in Fig. 1. For precipitation, the variation in GSS values was only slightly larger in Core than in S-Phys, and the control member usually had the highest skill at most times for both ensemble subsets (red curve). Although the two CIs shown in Fig. 1a have a very small amount of overlap at most times, the more robust pairwise difference test applied to these results (not shown) found the control member to have significantly higher GSSs at all times after hour 10 compared to almost all other members. Results for a threshold of ≥2.54 mm were similar but with values roughly 30%–40% lower at all times (not shown). Ideally, each member of an ensemble should be equally likely to verify, so there should be very little variation in a metric like GSS. However, for CREF, the variation in scores was noticeably larger in Core, and this increased difference comes about by having several members that are performing much more poorly than any member of S-Phys. Throughout the forecast period, members with P3 microphysics tended to have significantly lower GSS values than members employing other microphysics schemes. As with precipitation, the control member usually had the highest skill, significantly higher than all P3 and Morrison members at most times after hour 9 according to pairwise difference tests (not shown). In 2017 (not shown), the differences in GSS values by member for CREF were greatly reduced, and the values themselves were comparable to those of the majority of members in 2016 whose values were clustered just below that of the control run.

Frequency bias for 1-h precipitation thresholds of ≥2.54 and ≥6.35 mm (not shown) showed a more noticeable increase in score variations in the Core ensemble, along with many members of both ensembles often having too large of areal coverage at those thresholds. The same trends occurred in both 2016 and 2017. Frequency bias for CREF clearly indicated more variation in the metric among the members of Core than S-Phys, with additional differences between the 2016 (Fig. 2) and 2017 output (Fig. 3). In 2016, for the ≥20-dBZ threshold, all P3 members often significantly underpredicted areal coverage (frequency bias less than one) while most other members of Core significantly overpredicted at almost all times (except those using the Morrison microphysics), with frequency bias values greater than one. The control member nearly always exhibited the greatest overprediction, and it was often significantly larger than that of other members. At this same threshold, frequency bias values of all S-Phys members were more consistent with the control member. For a ≥40-dBZ threshold in 2016, the MY members had a very large overprediction, significantly larger than all other members at most times, and behaved notably differently from the other members, many of which exhibited an underprediction at most times. The tendency for the MY scheme to produce too high of reflectivity values within too broad convective cores had been noted

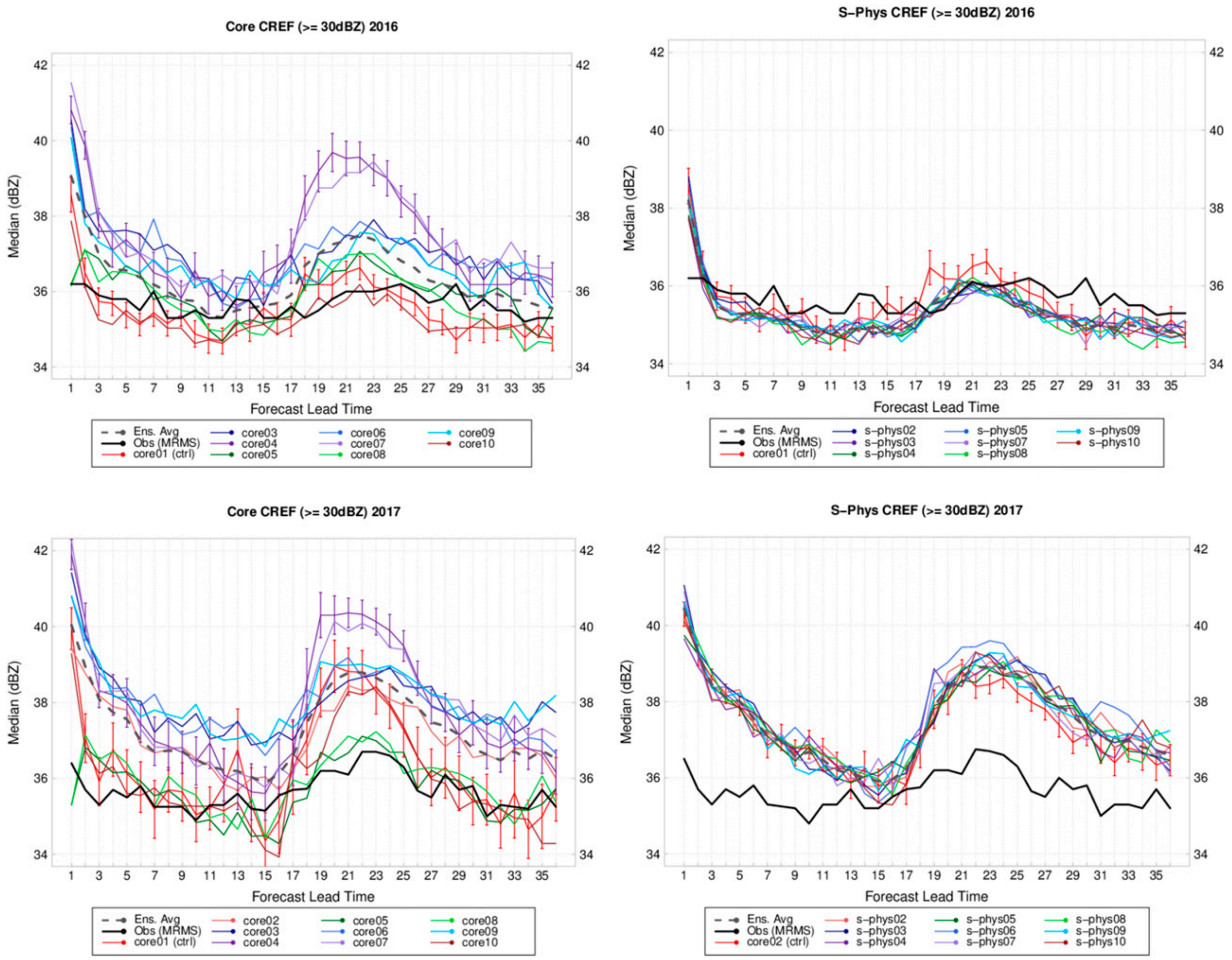


FIG. 5. Median CREF value (dBZ) for all members of (left) Core and (right) S-Phys for (top) 2016 and (bottom) 2017 for the ≥30-dBZ threshold. Colors of curves representing different members follow same notation as in Fig. 1 for 2016 and Fig. 3 for 2017, with observations in black and the ensemble average represented by a dark gray dashed line. The vertical bars represent 95% CIs for selected curves (core01 in all, with core04 added in left panels).

in Morrison et al. 2015), and attributed to too much graupel and excessive size sorting of hail. A clear diurnal signal in all members can be seen in the frequency bias values for the ≥40-dBZ threshold with higher values noted during the afternoon/evening hours (i.e., during typical convectively active periods) than at other times. This signal was also present in the precipitation fields (not shown) but does not show up for the lower CREF threshold of ≥20 dBZ (Fig. 2). In general, S-Phys members as a whole were more often closer to a value of 1 (unbiased) than Core members.

In 2017, Core again had a much larger variation in frequency bias values among members than S-Phys, and for ≥20 dBZ (Fig. 3), with statistically significant differences between the MY and several Thompson and P3 members at most times. In these cases, the control member no longer typically exhibited the highest overprediction, implying a reduction in areal coverage in the members due to the use of the different land surface and PBL schemes in 2017. The P3 members also did not have the underprediction problem that was present in 2016, as a change had been made to the scheme to improve the areal coverage of stratiform rain based upon 2016 SFE observations of a dry bias (J. Milbrandt, Environment and Climate Change Canada, 2019, personal communication). In addition, all S-Phys members had reduced frequency bias values compared to 2016, though the temporal behavior looked similar. For ≥40 dBZ, the MY members still had a significantly high overprediction, significantly higher than many members at most times, and were joined by the P3 members. A large change was evident in S-Phys at ≥40 dBZ for 2017 where a significant overprediction was present in most members at several lead times.

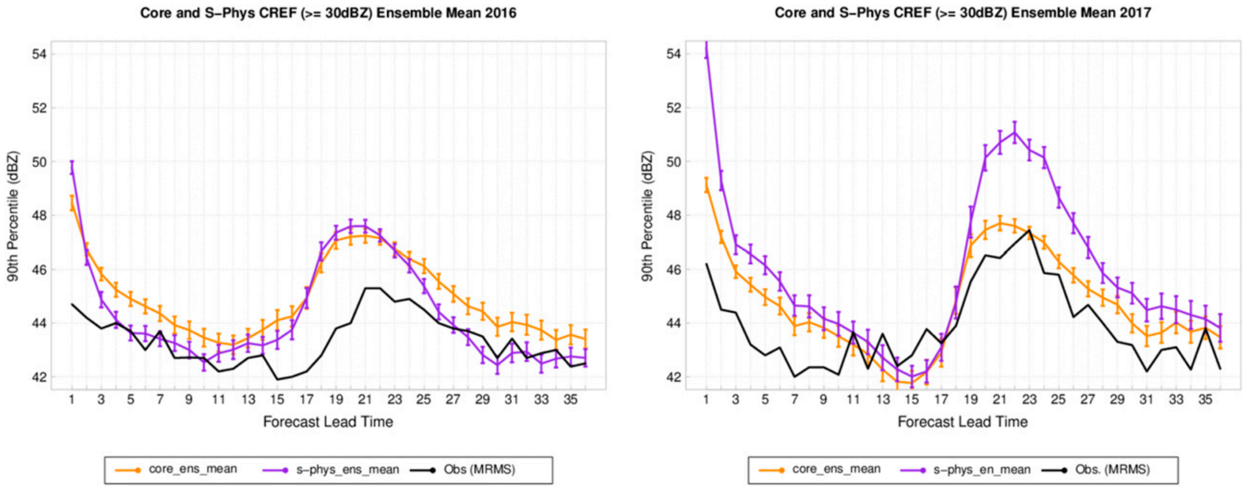


FIG. 6. The 90th percentile of composite reflectivity values for the ≥30-dBZ threshold averaged among the ensemble members for the Core (orange) and S-Phys (purple) ensembles in (left) 2016 and (right) 2017. Observed value shown in black. The vertical bars represent 95% CIs.

b. MODE verification

Several different object attributes were compared using MODE for the two ensembles in both years. The MODE settings are listed here with the select settings defined in parentheses. A convolution threshold (conv_thresh) of ≥2.54 mm was applied to identify the forecast and observation objects and a circular smoother (conv_radius) of 5 grid points was used. In addition, the following fuzzy engine weights were used: centroid distance (2), boundary distance (4), the distance between object orientation angles (1), the ratio of the object area (1), and the intersection area ratio (1). For 1-h precipitation, the median value based on a threshold of ≥2.54 mm to define the object areas (Fig. 4) shows greater variability for Core than S-Phys. Of note, for 1-h precipitation, almost all members in both ensembles in 2016 are greater than the observations (black curve), with the majority significantly so, except in the period of roughly 17–23 h in Core and 17–30 h in S-Phys. The same is true in 2017 (not shown). In both years, the increased intermember variability in Core does not translate into a better representation of the observations within the envelope. For heavier precipitation, the 90th percentile values indicate similar behavior in both years (not shown). When averaging members for each subensemble together (dashed dark gray line in Fig. 4), the value for both ensembles was also often significantly too high compared to observations, with Core at most times showing an increase in error by a magnitude of roughly 0.1 mm. Again, the 90th percentile averaged values (not shown) were too high at most times compared to the observations, with Core continuing to have larger error. As can be inferred from Fig. 4, the highest bias was present in both ensembles in the late night and morning hours when convection is normally weakening. During the period where severe convection is most likely, roughly between forecast hours 21 and 30, the S-Phys members did not show significant bias, while both ensembles had a small negative bias around the time of a diurnal minimum in observed precipitation (midday, roughly forecast hours 17–21).

The median CREF values for each member of both ensembles are displayed in Fig. 5. The increased variation in Core is apparent in these plots for both 2016 and 2017, and unlike with 1-h precipitation, the increased member variability results in a much better representation of the observations within the envelope of Core. While on the other hand, in both years, the observations fell outside any member prediction in S-Phys nearly all of the time. It should be noted, however, that although the observations were better captured in Core, some of its members significantly overestimated the reflectivity values. It can be seen that the median values were highest in the MY members during the afternoon hours when convection typically initiates. At these times, the MY values were significantly higher than all other members. Of note, for S-Phys, reflectivity values by member were often less than observations in 2016, significantly so about half the time for all members, but almost always significantly greater than observations in 2017. This might indicate that the use of the MYNN PBL scheme and the RUC land surface scheme resulted in more intense

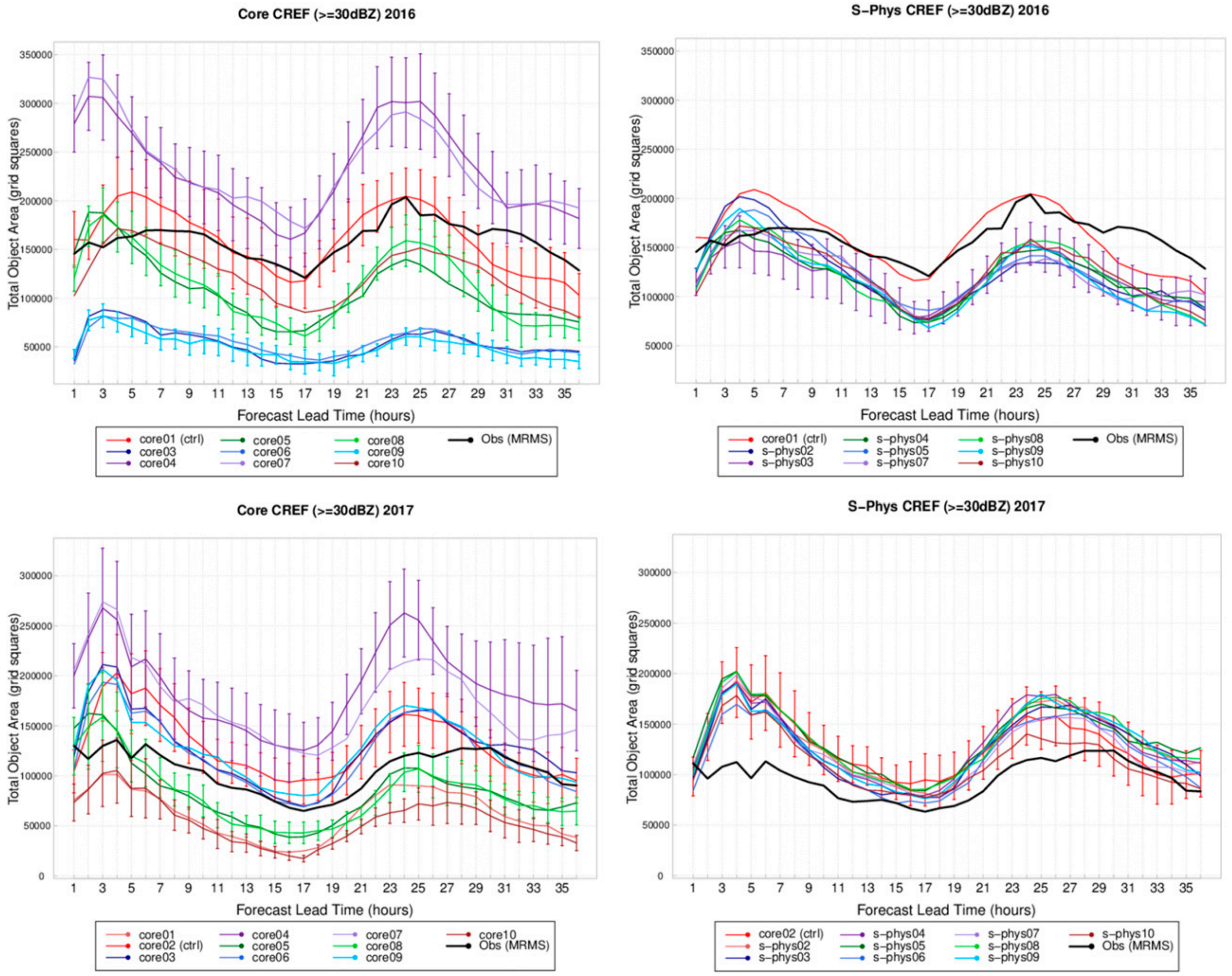


FIG. 7. Areas (grid squares) within the MODE objects for (left) Core and (right) S-Phys members in (top) 2016 and (bottom) 2017 for the ≥30-dBZ threshold. Colors of curves representing different members follow same notation as in Fig. 1 for 2016 and Fig. 3 for 2017, with observations in black. The vertical bars represent 95% CIs for selected curves (core01, core04, core08, and core09 in top left, core02, core04, core08, and core10 in bottom left, s-phys03 in top right, and core02 in bottom right).

reflectivity, but further work is needed to quantify this impact fully, as differences in predominant storm type between the two years could also result in some changes. A clustering of reflectivity values was also noted for the 90th percentile values (not shown) where the S-Phys members in both years were usually significantly too high compared to observations.

Similar behavior is apparent in the averaged values for the 90th percentile for CREF in Core (Fig. 6) for 2016, with the average being significantly too high compared to observations. However, in 2017, unlike for median CREF (Fig. 5), the average for Core agreed better with observations at most times after forecast hour 10, although it was still significantly higher at many lead times after forecast hour 17. For S-Phys, the 90th percentile was significantly too high during the daytime hours but close to the observations overnight in both years. Both ensembles showed a peak in values during the afternoon, around forecast hours 20–22, which was also observed, although usually the ensembles were too intense with the peak. Interestingly, observations did not show as pronounced a peak in the median values during the afternoon (Fig. 5), whereas the ensembles do depict a strong peak, thus increasing errors at that time, except for S-Phys in 2016.

Total areas within all MODE objects for each year for CREF (≥30 dBZ) are shown in Fig. 7. As would be expected, these results should be somewhat similar to the traditional metric of frequency bias. Much greater variability existed in the total areas of objects in the Core ensemble compared to S-Phys. In 2016, the control member (in red) lay closest to the observed value (black) at nearly all times and was not

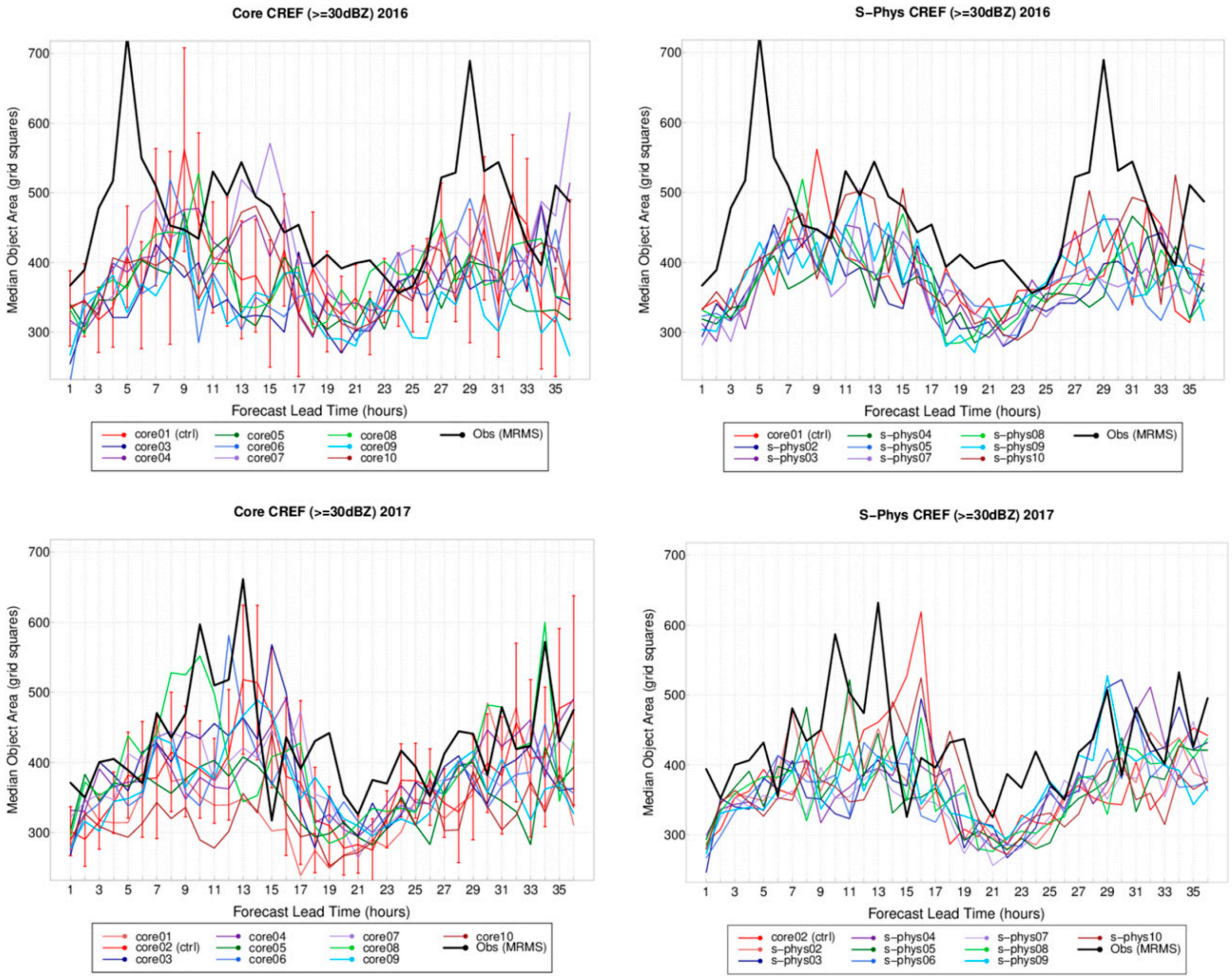


FIG. 8. Median area (grid squares) within the MODE objects for (left) Core and (right) S-Phys members in (top) 2016 and (bottom) 2017 for the ≥30-dBZ threshold. Colors of curves representing different members follow same notation as in Fig. 1 for 2016 and Fig. 3 for 2017, with observations in black. The vertical bars represent 95% CIs for core01 in left panels.

significantly different from observations at most times. In 2017, this was not the case as often, implying a worsening in the forecasts of areal coverage when the PBL scheme and land surface schemes were switched to MYNN and RUC, respectively. Overall, the Core ensemble did a better job of capturing the observed value within the envelope of members. This is especially true in 2017 when Core always had the observations within the envelope, usually around the middle of the envelope. For S-Phys, the observations were frequently outside the envelope of nearly all members, which underestimated areal coverage significantly (except the control member) in 2016 and frequently significantly overestimated the areal coverage in 2017. Again, this implies a potentially large impact from the change made in the PBL and land surface scheme in 2017. It should be noted in both years that a distinct clustering by microphysics scheme occurs in Core with curves rarely crossing each other and the values being significantly different at most times. This suggests that different physics combinations have very systematic differences in the amount of reflectivity ≥ 30 dBZ with limited variability over time (i.e., one member will always have broader areas of reflectivity; another member will always have much less). Such behavior again is concerning and provides further evidence of deficiencies when ensemble membership is based on a multiphysics approach as each member would not be equally likely to verify.

An analysis of the median area across all objects defined using a ≥30-dBZ CREF threshold (Fig. 8) looks much different from Fig. 7. The area of observed objects peaked around 0500 and 1100–1300 UTC, likely associated with large nocturnal MCSs, with a

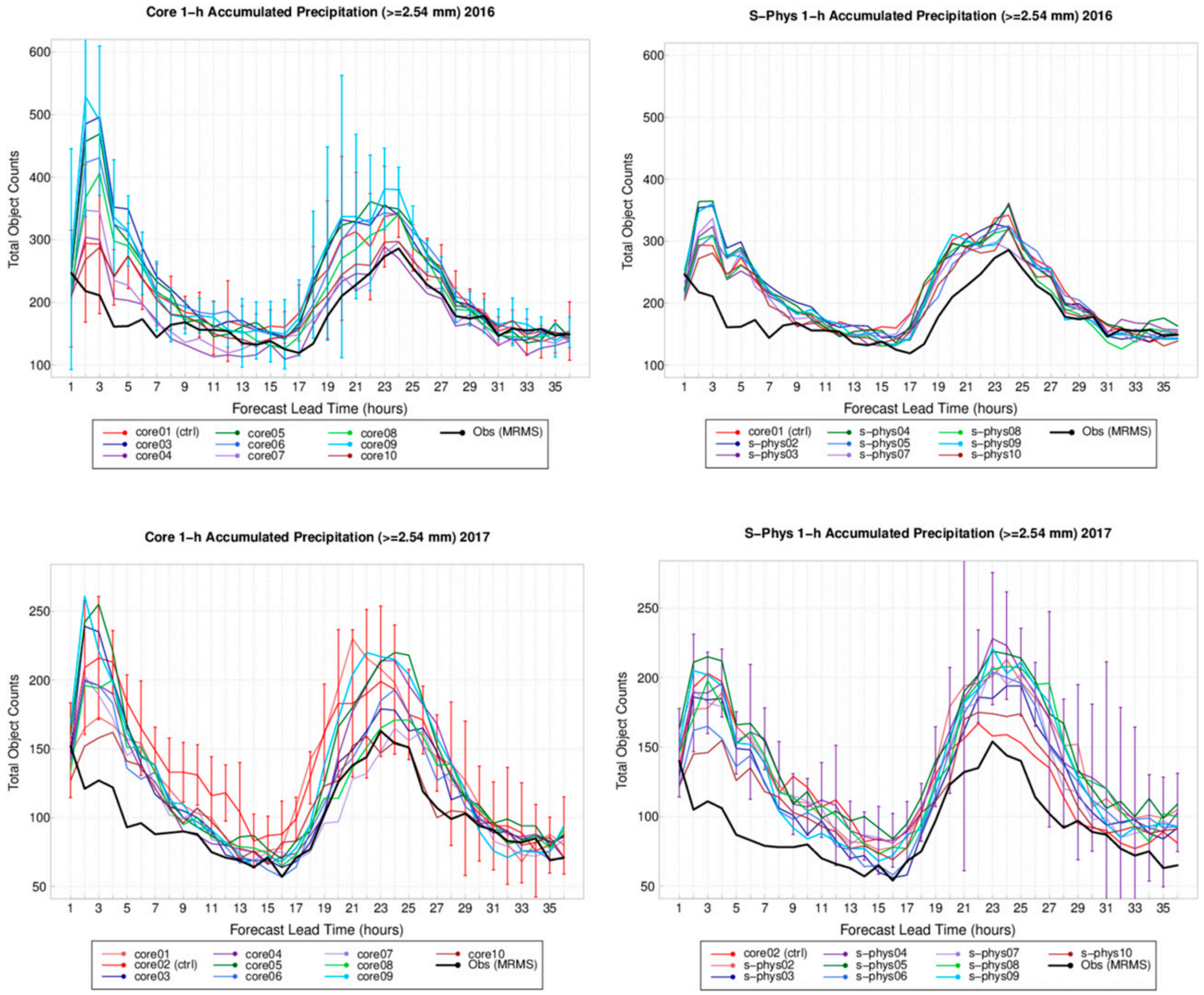


FIG. 9. Total count of MODE 1-h precipitation objects for (left) Core and (right) S-Phys members in (top) 2016 and (bottom) 2017 for the ≥2.54 mm threshold. Colors of curves representing different members follow same notation as in Fig. 1 for 2016 and Fig. 3 for 2017, with observations in black. The vertical bars represent 95% CIs for selected curves (core01 and core09 in upper left, core01 in lower left, and s-phys07 in lower right).

smaller peak in the afternoon. Objects defined using CREF were much smaller in all ensemble members around the time of the observed 0500 UTC peak, but the members did show a peak around the 1100–1300 UTC time period, albeit in most members the area was still significantly less than that observed. Unlike many other MODE attributes, variation in the median object areas seems comparable in S-Phys to Core.

Objects defined using a precipitation threshold of ≥2.54 mm (not shown) varied greatly in size from hour to hour during the first 18 h of the forecast with no substantial differences in the two ensembles; however, in the final 18 h of the forecast, the S-Phys members generally tended to produce objects smaller than those in Core and those observed.

Although areas differed among the Core members, the differences were far less than for CREF, with a roughly 60%–100% variation from the median in Core for CREF but only a 10%–20% variation in Core for 1-h precipitation. For S-Phys at most times, variations were roughly only 10% for both CREF and 1-h precipitation (the one exception was in 2016 where the control run deviated more from the other 8 members). Nonetheless, Core still did a better job of capturing the observations within its envelope.

For 1-h precipitation objects defined by a threshold of ≥2.54 mm, the number of objects identified in all members of both ensembles exceeded the number of observed objects at a majority of lead times during 2016 and 2017 (Fig. 9). This was especially true during the afternoon

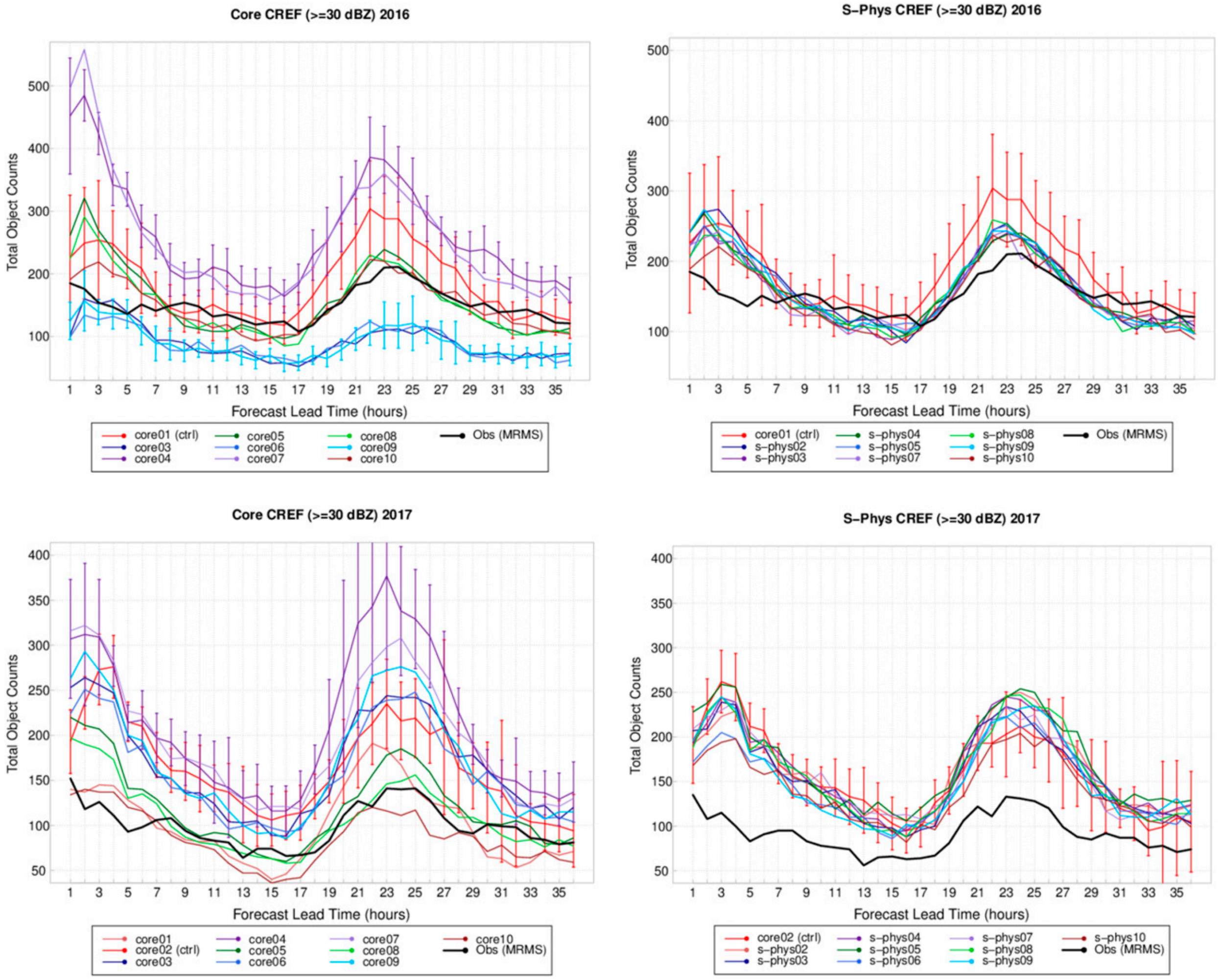


FIG. 10. Total count of MODE CREF objects for (left) Core and (right) S-Phys members in (top) 2016 and (bottom) 2017 for the ≥30-dBZ threshold. Colors of curves representing different members follow same notation as in Fig. 1 for 2016 and Fig. 3 for 2017, with observations in black. The vertical bars represent 95% CIs for selected curves (core01, core04, and core09 in top left, core01 in top right, core02 and core04 in bottom left, and core02 in bottom right).

and evening hours, when the differences were statistically significant for many members. In 2016, several Core members had substantially higher numbers of objects than any S-Phys members during convectively active times of day, although the large CI shown for core09 in Fig. 9 indicates the differences were not statistically significant. In general Core had a larger variation of counts of objects among members than in S-Phys. However, in 2017, the counts by member were much more similar between the two subensembles. Of note, a time offset existed in the ensembles, evidence that the models are generally a few hours too early with convective initiation.

For CREF, S-Phys members had smaller variability in their counts of objects, and showed counts similar to observations during certain portions of the day in 2016 with significantly too many objects forecast in the afternoon/evening hours and throughout much of the period in 2017 (Fig. 10). Core members separated into three clusters based on the microphysics scheme used, particularly in 2016, with MY members having roughly twice as many objects as observed, P3 members often only having half of the observed numbers, and Morrison members roughly matching the observed numbers after the first 6 h of the forecast period. The differences for MY and P3 were statistically significant at most times. Thus, much larger variability existed in the Core ensemble object counts than in S-Phys, and as might be expected, the observations were better contained within the envelope of its solutions.

The general displacement behavior for the 1-h precipitation objects were examined using the centroid attribute (center of mass) derived from MODE

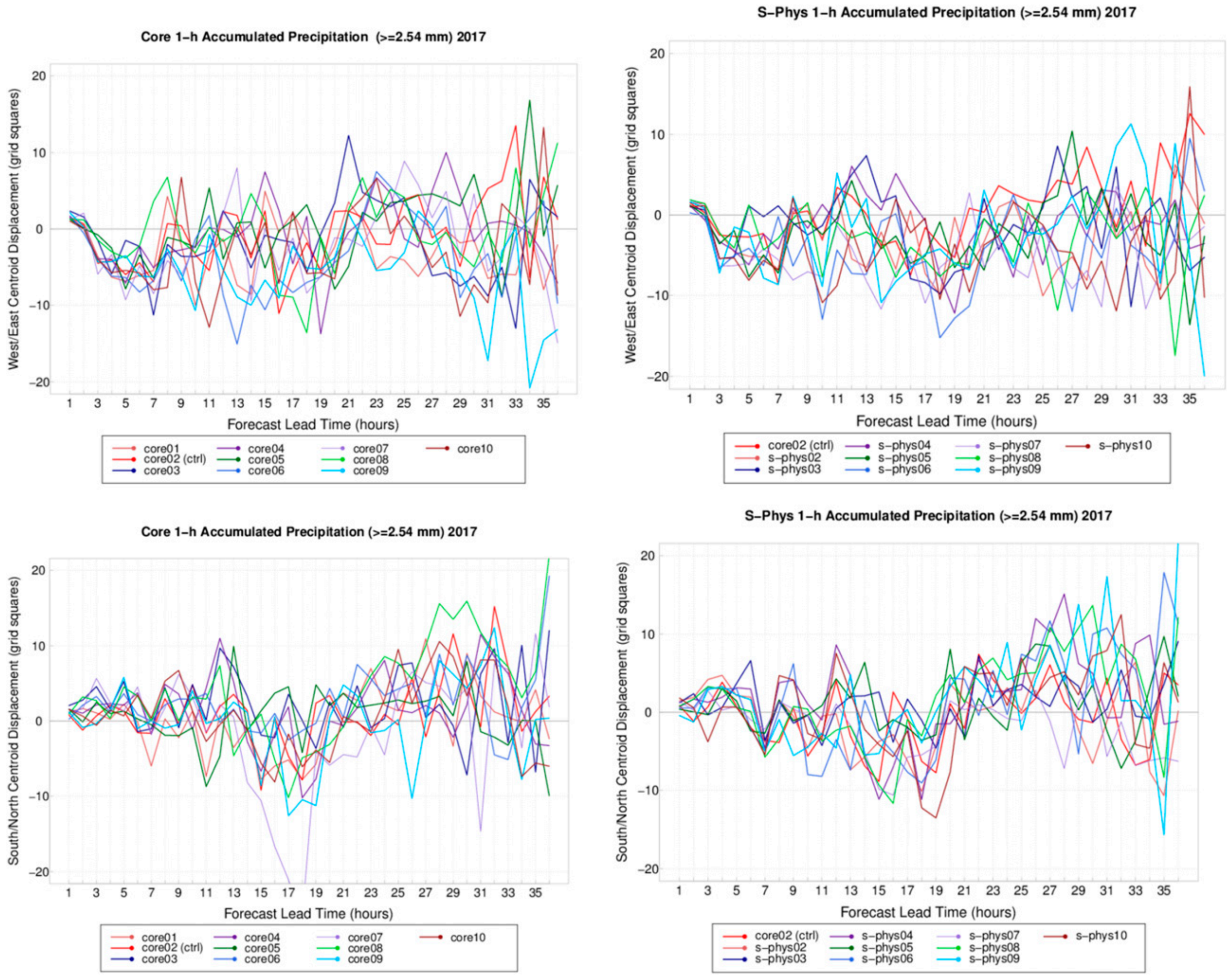
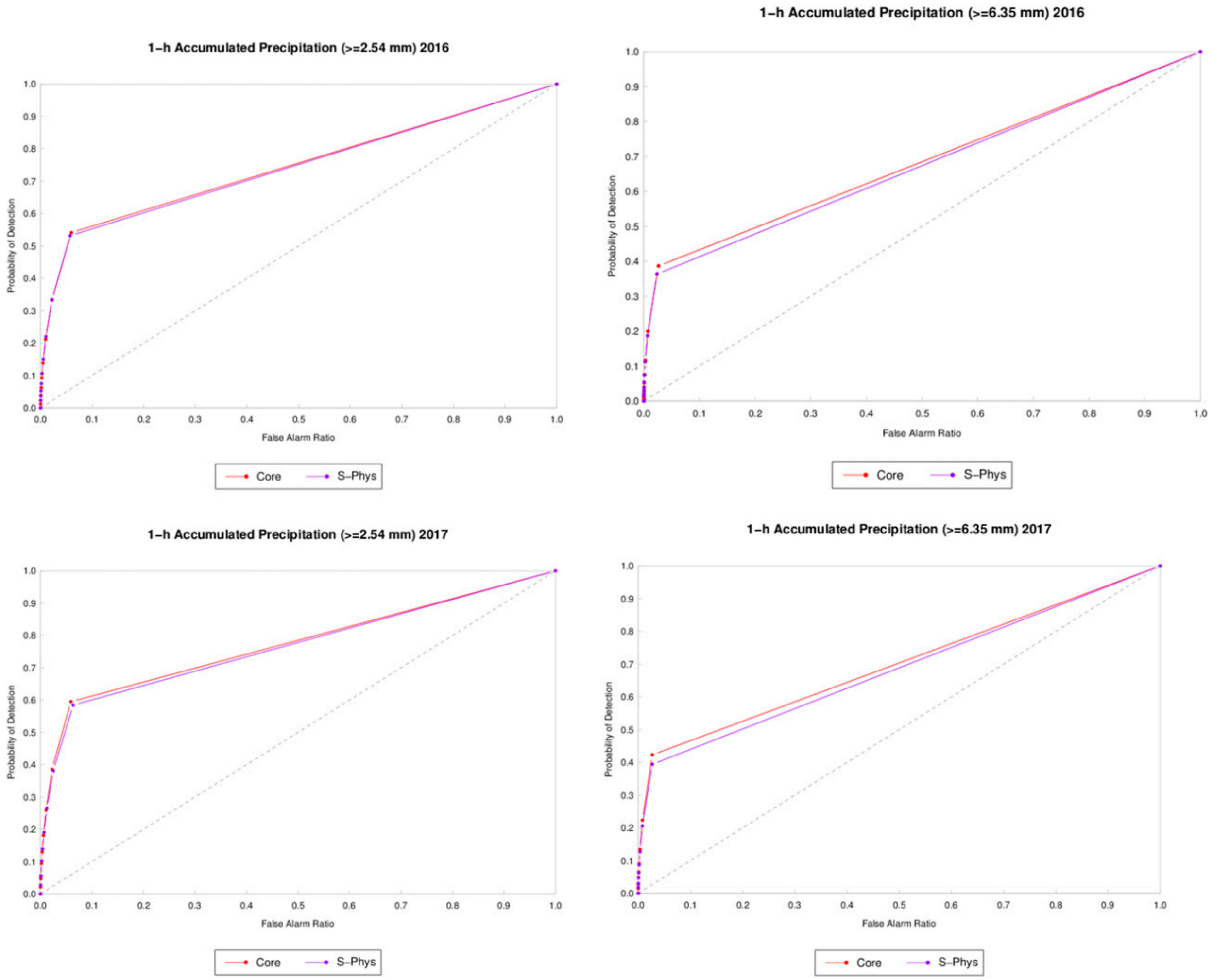


FIG. 11. Centroid displacement for 2017 in the (top) west–east direction and (bottom) south–north direction for (left) Core and (right) S-Phys members for 1-h precipitation objects for the ≥2.54-mm threshold. Colors of curves representing different members follow same notation as Fig. 3, with observations in black.

and calculating the centroid distance between the matched forecast and the observed objects (Fig. 11). For the west–east aspect, a negative (positive) value indicates a westerly (easterly) displacement and for the south–north aspect a negative (positive) value indicates a southerly (northerly) displacement. A majority of all examined ensemble members displayed westward displacement throughout the forecast period for 2017, although the trend was not statistically significant (thus no CIs are shown in Fig. 11). Early in the forecast period, members of both subensembles showed a slight easterly displacement followed by a sharp change toward a westerly displacement. This is potentially due to the fact that ongoing convection at the time of the 0000 UTC initialization was not well assimilated in the model and lacks sufficient cold pools to translate the storms eastward. Squitieri and Gallus (2019, manuscript submitted to *Wea. Forecasting*) found that cold pools were smaller, more shallow, and weaker in 3-km horizontal grid spacing simulations of MCSs than in 1-km simulations, while Verrelle et al. (2015) also found that cold pools were smaller in scale in coarser grid simulations than in finer ones. More investigation is required to confirm that cold pool deficiencies were present in the current sample of events. In general, the 2016 counterparts had somewhat less westward displacement and slightly less member variability (not shown).

In terms of the north–south displacement (Fig. 11), ensembles start out together and variance and error gradually increases to the south with more variance during the period of waning convection in morning hours. Then as convection intensifies in the afternoon, the displacement trend reverses, decreasing southward displacement trends


1–h Accumulated Precipitation (>=2.54 mm) 2016

1–h Accumulated Precipitation (>=6.35 mm) 2016

1–h Accumulated Precipitation (>=2.54 mm) 2017

1–h Accumulated Precipitation (>=6.35 mm) 2017

FIG. 12. ROC curves for Core (red) and S-Phys (purple) in (top) 2016 and (bottom) 2017 for (left) ≥2.54- and (right) ≥6.35-mm 1-h precipitation thresholds.

northward and increased variance among members continues into the evening. This shift to a northerly bias was more clearly noted in S-Phys. While the shift to the north persisted through the later forecast period in 2017, it was more transient in 2016 and began to turn southward again in the last few forecast hours (not shown). The north–south displacements were generally not statistically significant.

Overall, displacement in CREF had the same trend as the results described for 1-h accumulated precipitation (not shown).

c. *Traditional ensemble verification*

ROC curves, areas under the curves, reliability diagrams, and Brier scores were examined for both ensembles, and generally indicated only a slight advantage at best for the Core ensemble. ROC curves for both years can be seen in Fig. 12 for two 1-h rainfall thresholds. In 2016, the two curves were nearly identical for ≥2.54 mm, while Core had a slight advantage for ≥6.35 mm. Although not shown, Core had a bigger advantage in area under the ROC curve at most times for ≥0.254 mm. The improvement of Core over S-Phys was slightly larger in 2017 for both thresholds. In both years, skill (area under the ROC curve > 0.7) existed at a majority of times for the ≥0.254- and ≥2.54-mm thresholds. Skill was only present for the first 6–12 h of the forecast for the ≥6.35-mm threshold (not shown). ROC areas were generally 0.01–0.02 greater for 3-h precipitation (not shown) with skillful forecasts extending to later lead times for the 6.35-mm threshold.

Reliability diagrams for 1-h precipitation suggested a similar small advantage for the Core ensemble in 2016 (Fig. 13), but both ensembles overestimated the probabilities except for 0%, with curves lying well to the right of the diagonal lines. Skill relative to

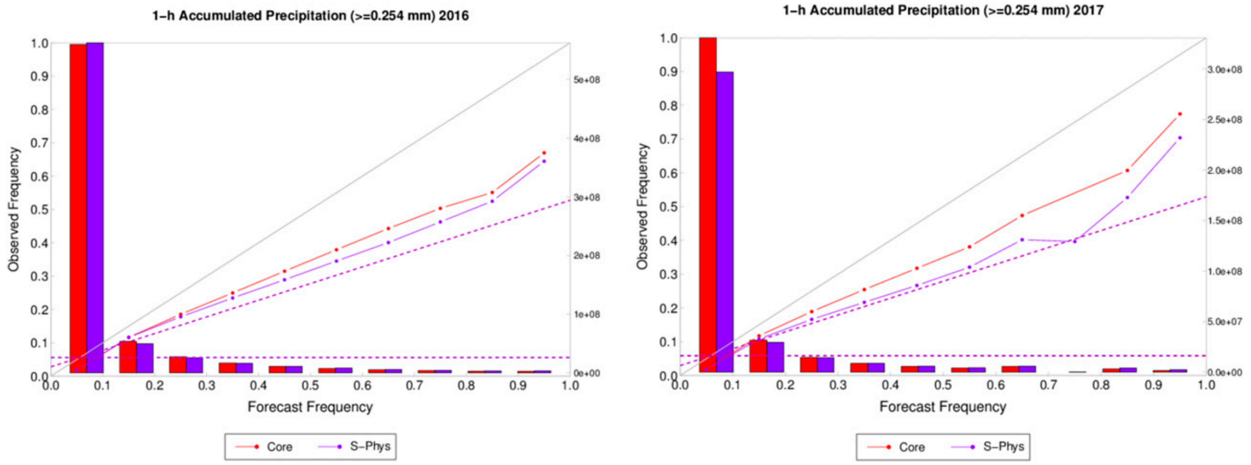


FIG. 13. Reliability curves for Core (red) and S-Phys (purple) for a precipitation threshold of ≥0.254 mm in 1 h for (left) 2016 and (right) 2017. Histogram shows the counts for each forecasted probability. Black solid diagonal line represents perfect reliability; purple dashed diagonal line is the no skill line. Horizontal purple dashed line represents climatology.

climatology only existed in 2016 for both ensembles for 1-h precipitation above ≥0.254 mm. Some skill was present for 3-h precipitation at the 2.54-mm threshold (not shown). The Core ensemble performed better in 2017 and showed some skill for 1-h precipitation even at the ≥2.54-mm threshold. For 3-h precipitation, the Core ensemble was relatively reliable with its curve close to the diagonal line (not shown). The difference in performance between Core and S-Phys was much greater in 2017 than in 2016, perhaps suggesting again that the change in PBL scheme and land surface scheme harmed the S-Phys ensemble in 2017.

Brier Scores for both 1-h precipitation and CREF showed the same behavior as that found with the ROC curves and reliability diagrams, with limited differences between the two models (not shown).

d. Convective initiation verification

The investigation of convective initiation found there was less variation in the location of the initiation in S-Phys than in Core, but also smaller peak errors on average among the members of S-Phys for the 10 cases (Fig. 14). Both ensembles had the observed location within the envelope of member solutions in 6 of the 10 cases. Although this subset of 10 events, chosen based on relatively pristine daytime initiation of substantial convective systems, represents only a very small subsample of all objects during 2016, some similarities can be seen with the displacement errors shown for all objects in 2017 (Fig. 11) during the afternoon hours. Specifically, the difference in spread in the latitudinal direction for these convective initiation cases between the two subensembles (Fig. 14) was more than the difference in longitudinal spread. Figure 11 suggests overall a slightly greater variation in north–south displacement errors among the Core members than the S-Phys members during the afternoon.

4. Discussion and summary

Two CLUE subensembles were examined in detail to study the impact of including mixed physics in an ensemble that already used perturbed IC/LBCs. Comparisons were made using 22 cases of 1- and 3-h precipitation and 17 cases of CREF in 2016, as well as 12 cases of both precipitation and CREF from 2017 CLUE output. Multiple verification metrics were examined.

In most cases, the mixed physics ensemble (Core) had noticeably more variation in verification scores than the Single physics (S-Phys) ensemble. Differences in values were larger when evaluating CREF as opposed to precipitation, with much more variation, often statistically significant, showing up in the reflectivity fields. This is likely because small changes in assumed hydrometeor size distributions can cause large changes in model reflectivity fields but negligible changes in precipitation rates (J. Milbrandt, Environment and Climate Change Canada, 2019, personal communication). In most cases, but not all, the increased variability in Core better captured the observed value, and S-Phys appeared to be substantially underdispersive at most times. The ensemble average value agreed better with observations for Core compared to S-Phys. However, especially for reflectivity, this better average value came about because members like those that used MY microphysics, which had large positive errors in intensity and areal coverage, tended to balance negative errors found in many of

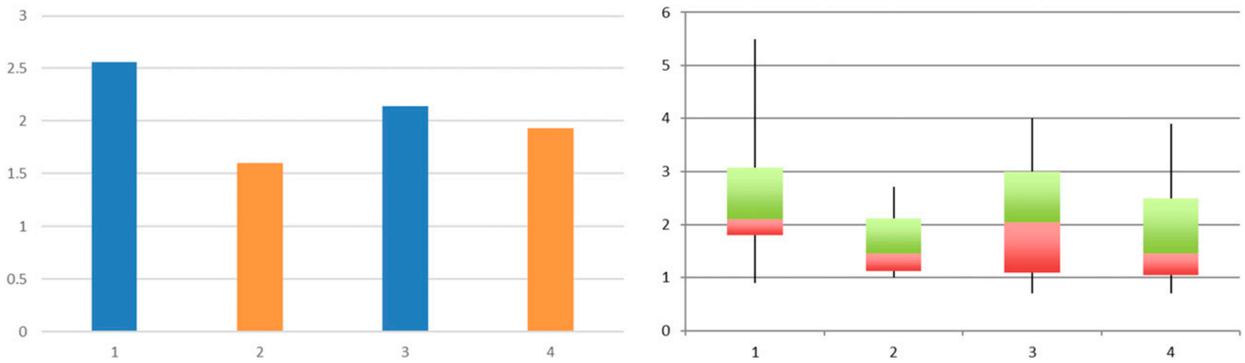


FIG. 14. (left) Variation in average spread among the 9 ensemble members for the 10 cases for Core (blue) and S-Phys (orange), with latitude spread in boxes 1 and 2 and longitude spread in boxes 3 and 4, and (right) whisker plots of maximum errors in latitude (°; leftmost 2 bars) and longitude (°; rightmost 2 bars) among the 9 members for the 10 cases. Boxes 1 and 3 are for Core, and boxes 2 and 4 are for S-Phys.

the other configurations. A summing of the errors from individual members would reveal Core to be worse, thus, further evidence of deficiencies when ensemble membership is based on a multiphysics approach. Traditional ensemble measures gave a slight advantage to the mixed physics ensemble, but suggested very little skill for 1-h precipitation. More skill was present for 3-h precipitation.

A similar increase in variability was shown in an evaluation of 10 cases of pristine convective initiation, new convection forming relatively far from existing convection, from the 2016 sample of cases. However, despite the increased spread in latitude and longitude positioning of initiation in Core, both ensembles correctly captured the observed location within their envelope of solutions in 60% of the cases. Thus, the performance of the two ensembles might be regarded as equal.

The results from this study raise several questions. First, is the increased variability in Core a benefit to forecasters? Second, do the slight advantages shown for Core in some skill measures justify the identified issues that are associated with mixed-physics ensembles? Third, with such strong biases present when some microphysics schemes are used, would a better designed mixed physics ensemble that uses different microphysics schemes that have less extreme biases, or bias corrections made to the schemes used here, result in a more obvious improvement in skill over the single physics ensemble? Future work should explore the impact of bias corrections, particularly to the CREF values, and examine the performance of the two ensembles for other variables that are used to provide guidance to severe weather forecasters, such as updraft helicity and peak surface wind. In addition, future work should explore the use of a stochastically perturbed single physics ensemble as a means to reap some of the benefits associated with a mixed physics ensemble (e.g., increased spread) while avoiding problems associated with higher costs of maintaining multiple physics packages.

Acknowledgments. The authors thank the groups that contributed to the CLUE ensemble (the Center for Analysis and Prediction of Storms, the National Severe Storms Laboratory, the National Center for Atmospheric Research, the University of North Dakota, and the Earth System Research Laboratory/Global Systems Division of NOAA). Tressa Fowler assisted with the statistical significance tests. In addition, the constructive comments of three anonymous reviewers were appreciated. Support for this project was provided by the Developmental Testbed Center (DTC). The DTC Visitor Program is funded by the National Oceanic and Atmospheric Administration, the National Center for Atmospheric Research, and the National Science Foundation. In addition, partial support for the research was provided by NSF Grant AGS1624947.

REFERENCES

Benjamin, S. G., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Wea. Rev.*, **144**, 1669–1694, https://doi.org/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, https://doi.org/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, https://doi.org/10.1175/MWR-D-14-00091.1.

Buizza, R., M. Milleer, and T. N. Palmer, 2007: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **125**, 2887–2908, https://doi.org/10.1002/qj.49712556006.

Bullock, R., and Coauthors, 2017: Model Evaluation Tools version 6.1 (METv6.1): User's guide. Developmental Testbed Center, 400 pp., https://dtcenter.org/met/users/docs/users_guide/MET_Users_Guide_v6.1.pdf.

Clark, A. J., and Coauthors, 2012: An overview of the 2010 Hazardous Weather Testbed Experimental Forecast 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.

Davis, C., B. Brown, and R. Bullock, 2006a: Object-based verification of precipitation forecasts. Part I: Methods and application to mesoscale rain areas. *Mon. Wea. Rev.*, **134**, 1772–1784, https://doi.org/10.1175/MWR3145.1.

——, ——, and ——, 2006b: Object-based verification of precipitation forecasts. Part II: Application to convective rain systems. *Mon. Wea. Rev.*, **134**, 1785–1795, https://doi.org/10.1175/MWR3146.1.

——, ——, ——, and J. Halley-Gotway, 2009: The Method for Object-based Diagnostic Evaluation (MODE) applied to numerical forecasts from the 2005 NSSL/SPC Spring Program. *Wea. Forecasting*, **24**, 1252–1267, https://doi.org/10.1175/2009WAF2222241.1.

Du, J., G. DiMego, S. Tracton, and B. Zhou, 2003: NCEP short-range ensemble forecasting (SREF) system: Multi-IC, multi-model and multi-physics approach. *Research Activities in Atmospheric and Oceanic Modelling*, J. Cote, Ed., Rep. 33, CAS/JSC Working Group Numerical Experimentation (WGNE), WMO/TD-1161, 5.09–5.10.

Gallo, B. T., A. J. Clark, and S. R. Dembek, 2016: Forecasting tornadoes using convection-permitting ensembles. *Wea. Forecasting*, **31**, 273–295, https://doi.org/10.1175/WAF-D-15-0134.1.

Gallus, W. A., Jr., 2010: Application of object-oriented verification techniques to ensemble precipitation forecasts. *Wea. Forecasting*, **25**, 144–158, https://doi.org/10.1175/2009WAF2222274.1.

Gilleland, E., 2010: Confidence intervals for forecast verification. NCAR Tech. Note NCAR/TN-479+STR, 71 pp., https://doi.org/10.5065/D6WD3XJM.

Hacker, J. P., and Coauthors, 2011: The U.S. Air Force Weather Agency's mesoscale ensemble: Scientific description and performance results. *Tellus*, **63A**, 625–641, https://doi.org/10.1111/j.1600-0870.2010.00497.x.

Hong, S., Y. Noh, and J. Dudhia, 2006: A new vertical diffusion package with an explicit treatment of entrainment processes. *Mon. Wea. Rev.*, **134**, 2318–2341, https://doi.org/10.1175/MWR3199.1.

Houtekamer, P. L., and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.*, **126**, 796–811, https://doi.org/10.1175/1520-0493(1998)126<0796:DAUAEK>2.0.CO;2.

Hu, M., M. Xue, and K. Brewster, 2006: 3DVAR and cloud analysis with WSR-88D Level-II Data for the prediction of Fort Worth tornadic thunderstorms Part I: Cloud analysis and its impact. *Mon. Wea. Rev.*, **134**, 675–698, https://doi.org/10.1175/MWR3092.1.

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, https://doi.org/10.1175/1520-0493(1994)122<0927:TSMECM>2.0.CO;2.

Jankov, I., and Coauthors, 2017: A performance comparison between multiphysics and stochastic approaches within a North American RAP ensemble. *Mon. Wea. Rev.*, **145**, 1161–1179, https://doi.org/10.1175/MWR-D-16-0160.1.

Johnson, A., X. Wang, M. Xue, and F. Kong, 2011: Hierarchical cluster analysis of a convection-allowing ensemble during the Hazardous Weather Testbed 2009 Spring Experiment. Part II: Ensemble clustering over the whole experiment period. *Mon. Wea. Rev.*, **139**, 3694–3710, https://doi.org/10.1175/MWR-D-11-00016.1.

——, ——, J. R. Carley, L. J. Wicker, and C. Karstens, 2015: A comparison of multiscale GSI-based EnKF and 3DVar data assimilation using radar and conventional observations for midlatitude convective-scale precipitation forecasts. *Mon. Wea. Rev.*, **143**, 3087–3108, https://doi.org/10.1175/MWR-D-14-00345.1.

Milbrandt, J. A., M. K. Yau, J. Mailhot, and S. Belair, 2008: Simulation of an orographic precipitation event during IMPROVE-2. Part I: Evaluation of the control run using a triple-moment bulk microphysics scheme. *Mon. Wea. Rev.*, **136**, 3873–3893, https://doi.org/10.1175/2008MWR2197.1.

Mitchell, K. E., and Coauthors, 2001: The Community Noah Land Surface Model (LSM): User's guide (version 2.7.1). NCEP, 26 pp., https://ral.ucar.edu/sites/default/files/public/product-tool/unified-noah-lsm/Noah_LSM_USERGUIDE_2.7.1.pdf.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF Ensemble Prediction System: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73–119, https://doi.org/10.1002/qj.49712252905.

Morrison, H., and J. A. Milbrandt, 2015: Parameterization of cloud microphysics based on the prediction of bulk ice particle properties. Part I: Scheme description and idealized tests. *J. Atmos. Sci.*, **72**, 287–311, https://doi.org/10.1175/JAS-D-14-0065.1.

——, G. Thompson, and V. Tatarskii, 2009: Impact of cloud microphysics on the development of trailing stratiform precipitation in a simulated squall line: Comparison of one- and two-moment schemes. *Mon. Wea. Rev.*, **137**, 991–1007, https://doi.org/10.1175/2008MWR2556.1.

——, J. A. Milbrandt, G. H. Bryan, K. Ikeda, S. A. Tessendorf, and G. Thompson, 2015: Parameterization of cloud microphysics based on the prediction of bulk ice particle properties. Part II: Case study comparisons with observations and other schemes. *J. Atmos. Sci.*, **72**, 312–339, https://doi.org/10.1175/JAS-D-14-0066.1.

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.

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting*, **5**, 570–575, https://doi.org/10.1175/1520-0434(1990)005<0570:TCSIAA>2.0.CO;2.

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.

Smirnova, T. G., J. M. Brown, and S. G. Benjamin, 1997: Performance of different soil model configurations in simulating ground surface temperature and surface fluxes. *Mon. Wea. Rev.*, **125**, 1870–1884, https://doi.org/10.1175/1520-0493(1997)125<1870:PODSMC>2.0.CO;2.

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

Thompson, G., P. R. Field, W. D. Hall, and R. M. Rasmussen, 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.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317–2330, https://doi.org/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, https://doi.org/10.1175/1520-0434(1993)008<0379:OEPATN>2.0.CO;2.

Verrelle, A., D. Ricard, and C. Lac, 2015: Sensitivity of high-resolution idealized simulations of thunderstorms to horizontal resolution and turbulence parameterization. *Quart. J. Roy. Meteor. Soc.*, **141**, 433–448, https://doi.org/10.1002/qj.2363.

Wolff, J. K., M. Harrold, T. Fowler, J. Halley 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.

Xue, M., D.-H. Wang, J.-D. Gao, K. Brewster, and K. K. Droegemeier, 2003: The Advanced Regional Prediction System (ARPS), storm-scale numerical weather prediction and data assimilation. *Meteor. Atmos. Phys.*, **82**, 139–170, https://doi.org/10.1007/s00703-001-0595-6.

Zhang, J., and Coauthors, 2016: Multi-Radar Multi-Sensor (MRMS) quantitative precipitation estimation. *Bull. Amer. Meteor. Soc.*, **97**, 621–637, https://doi.org/10.1175/BAMS-D-14-00174.1.