DTC Visitor Program Final Report

Mesoscale Model Intercomparison and Analysis of Forecast Bias in Convective Structure Across Regional Convection-Allowing Summertime Forecasts

Mariusz Starzec and Gretchen Mullendore

Department of Atmospheric Sciences, University of North Dakota, Grand Forks, ND

Abstract

The skill of five local 3-km forecasts with varying microphysics and model versions are assessed by comparing the simulated reflectivity field against the radar reflectivity field as observed by two S-band weather radars. The experimental High-Resolution Rapid Refresh (HRRR; version 2) is also evaluated in order to compare skill between local “cold start” forecasts and operational forecasts with data assimilation. Model validation is performed by investigating the bulk number of convective objects and their areas with height. The distributions of reflectivity values within objects with height are investigated to evaluate if forecasts correctly simulated convective processes and storm depth.

WSM6 forecasts significantly over-predict the number of objects near the surface (especially weaker objects), which results in over-prediction of the area covered by convection. Nevertheless, while WSM6 over-predicts object counts, the intensity distribution of objects closely resembles observations. Thompson and HRRR forecasts perform better than WSM6 overall, but tend to over-intensity small convective cells. All model forecasts notably over-predicted the number of storms during the diurnal cycle. In some cases, model validation at one level reveals no major differences; however, by expanding the evaluation into the vertical, large differences are found that would not be noticed with a traditional two-dimensional evaluation. Forecasts typically contain a wider distribution of reflectivity values than observations and forecasts are frequently too intense, especially above the melting level. A maximum in area coverage related to the stratiform region is found in HRRR forecasts 1 to 3 km above the melting level, with no such signature in the observations.

1. Introduction

Convective forecast verification plays an important role in advancement of numerical weather prediction models; however, meaningful evaluations on convection-allowing scales still remain challenging. A form of verification that accounts for spatial differences is necessary for high-resolution forecasts as traditional skill scores have limited usefulness when spatial anomalies exist between forecasts and observations (e.g., Mass et al. 2002). It is known that the exact timing and location of convective initiation is very difficult to predict due to the chaotic nature of the boundary layer. The verification procedure should instead be more focused on evaluating convective characteristics, such as intensity, depth, size, and coverage, and less on the precise location. The timing of convective events, in particular, the timing of convective initiation is also important as an offset in timing may influence convective morphology and evolution.

Model forecasts have typically been evaluated by analyzing the simulated precipitation field against observed precipitation (e.g., Ebert and McBride 2000; Casati et al. 2004; Davis et al. 2006b; Clark et al. 2014). Knowing what biases are present in the accumulated precipitation fields has important implications on operational forecasting; however, it is also important to try and determine what processes lead to those biases. Since the precipitation field is temporally averaged, potential features and details are smoothed out. As a result of temporal smoothing, different forms of convection can produce a similar precipitation field while being dynamically and/or microphysically distinct from each other.

In this study, the reflectivity field at multiple heights is employed to evaluate forecasts and provide information on vertical structure. Unlike the accumulated precipitation field that results from convective processes, the reflectivity field is not temporally averaged and provides an instantaneous representation of convective processes. The reflectivity field at multiple levels also reveals how the convective structure evolves with height. Furthermore, comparing the differences between simulated reflectivity and observed reflectivity allows the convective systems to be directly compared and can lead to further insight as to why biases exist in the precipitation fields. The evaluation will primarily focus on object statistics with height and only a minor examination on convective timing is noted.

2. Data

a) Model Data

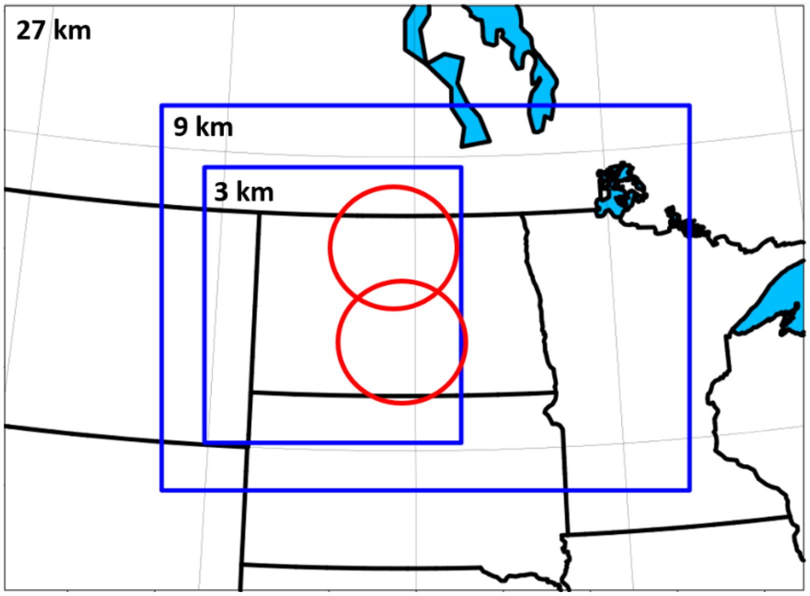

Figure 1. The grid structure used for local WRF forecasts. The red circles indicate a radius of 126 km from the Minot (top) and Bismarck (bottom) radars.

Five locally-run Weather Research and Forecasting (WRF) 3-km model realizations with differing microphysics and/or versions are evaluated and summarized in Table 1. All other model physics and parametrizations remain the same throughout the forecasts. The 3-km model domain where the analysis is performed is predominantly located over western North Dakota (ND; Fig. 1), embedded within a 9-km grid that is nested within a 27-km domain. Forecasts were initialized at 00 UTC (7 pm local time) daily starting 1 June 2015 through 1 October 2015 using the 40-km North American Mesoscale (NAM) model for initial and lateral boundary conditions. Forecasts were generated out to 48 hours from initialization. Forecasts were “cold-start;” however, to retain aerosol information for the Thompson “aerosol-aware” (TAA; Thompson and Eidhammer 2014) scheme, the aerosol concentrations were updated for daily initialization using the 24-hour forecast concentration from the previous day (i.e., 00 UTC valid time). Without this step, aerosol concentrations would not be retained but instead would default to the climatological values at each daily initialization. While the significance of retaining the aerosol concentration has not been evaluated, it is ideal that aerosol concentrations remain updated especially after undergoing processes that can greatly affect concentrations such as scavenging by rain. The greatest anticipated impact would likely be in timing of convective initiation and generation of rain via autoconversion.

Table 1. Local WRF model physics and configurations.

Local WRF Physics		
Version	**Microphysics**	**Planetary Boundary Layer:** YSU **Surface Layer:** MM5 Similarity **Land Surface:** Noah **Shortwave:** Dudhia **Longwave:** RRTM **Cumulus:** Kain-Fritsch (9 and 27 km only)
3.1.1	WSM6	
	Thompson	
3.7.0	WSM6	
	Thompson	
	Thompson “Aerosol-aware”	

In addition to the five WRF forecast setups, the experimental High Resolution Rapid Refresh (HRRR) model (version 2; Table 2) is also evaluated across the same period. Contrary to the local WRF forecasts, the HRRR model is initialized hourly and generates forecasts out to 27-hours for the continental United States. However, in addition to random data outages there were significant systematic data outages for the 18-27 hour forecast range in the evaluation period (i.e., June, July, and early September). Therefore, to retain data availability, two initializations of the HRRR are evaluated: 00 and 12 UTC. Both the 00 and 12 UTC initializations are evaluated for 12 hours starting at the 4-hr lead time out to a 15-hr lead time; therefore, the 00 UTC forecasts cover the valid times from 04 to 15 UTC and the 12 UTC forecasts are evaluated from 16 to 03 UTC. Furthermore, all model initializations are evaluated starting at the 4-hr lead time in order to avoid including potential spin-up biases.

Table 2. Experimental HRRR v2 model physics used during the summer of 2015.

Experimental HRRR v2 Model Physics	
Version	3.6.1
Microphysics	Thompson “Aerosol aware”
PBL	MYNN
Surface Layer	MYNN
Land Surface	RUC
Shortwave	RRTMG
Longwave	RRTMG

b) Radar Data

Radar observations taken by the WSR-88D radars located at Bismarck, ND (i.e., KBMX) and Minot, ND (i.e., KMBX) are made available by the National Centers for Environmental Information (NCEI). Radar data for each radar are interpolated to a uniform Cartesian grid with resolution of 1 km in the horizontal and 1 km in the vertical, extending from 2 km to 13 km high. Data are mapped out to 126 km from each radar (red circles; Fig. 1), which is the furthest distance where the bottom of the radar beam (assuming standard atmospheric refraction) still samples the primary 2 km analysis height (further discussed below). Only radar data nearest to the top of each hour are used in order to match the hourly output of the model forecasts. The data must be time-stamped to be within 5 minutes of every hour. When data are present from both radars for a specific hour, the overlapping data are spatially composited using a simple distance-weighted mean.

The reflectivity data contains many artifacts and non-meteorological echo, and is subjected to similar quality control methods as outlined in Tang et al. (2014). Tang et al. (2014) filter radar echo by performing quality control methods along radar radials. Since the radar data in this study is gridded, many procedures are not applicable; however, some methods in Tang et al. (2014) are adapted to the gridded field. Initially, all pixels with a correlation coefficient (ρ_{HV}) < 0.95 are removed unless they are identified as being possible hail cores or regions of melting. Hail cores are considered such if the reflectivity (Z_H) of a pixel is > 45 dBZ and 18 dBZ echo tops > 8.0 km (Tang et al. 2014). The melting level is found where pixels contain $\rho_{HV} > 0.7$ and $Z_H > 30$ dBZ in the lowest 5 km, with reflectivity data being present in at least three layers of the column stretching from the surface to 5 km. Following the ρ_{HV} filter, pixels containing very high values of differential reflectivity (Z_{DR}) > 5 dB coupled with low values of $Z_H < 10$ dBZ are removed, as these Z_{DR} signals are typically present from biological scatters. Lastly, Z_H filters are applied to data. Individual pixels are filtered out if the rate of change of Z_H in the lowest two heights is > 50 dBZ, Z_H of the lowest level pixel is > 30 dBZ with no echo above, or if the Z_H < 0 dBZ. After all Z_H filtering, a nearest neighborhood method is employed to remove clutter that contains high ρ_{HV} but was not removed by the ρ_{HV} filter. For a given pixel containing Z_H data, if more than half of the neighboring pixels have missing reflectivity data, the pixel is filtered (Tang et al. 2014). Remaining echoes comprised of three pixels (i.e., 3 km^2) or less are removed. After filtering, any pixel-sized gaps created by the initial ρ_{HV} filter that are surrounded by meteorological echo are returned to their original Z_H value (similar to Tang et at. 2014). While the quality of radar data is greatly improved with these quality control metrics, there still are occasional artifacts left over primarily in the overnight hours during times of radar beam ducting.

3. Evaluation of Model Forecasts

The model forecasts are evaluated by comparing the simulated reflectivity field to the observed reflectivity field at various heights ranging from 2 to 13 km, with 2 km being the primary evaluation height. Comparisons are performed every forecast valid hour, based on data availability. The simulated reflectivity field is calculated following Koch et al. (2005). The simulated radar reflectivity factor in the Koch et al. (2005) method is determined using assumptions about the distributions of hydrometeor size and shape for rain, snow, and graupel with fixed distribution intercept for WSM6 and variable intercepts for the Thompson-based schemes. While the 1-km height is ideal due to the proximity of echo to the surface and frequent use by forecasters, the 2-km height is chosen to expand the area of analysis. The 1-km height limited the area of analysis enough to produce detrimental effects on the verification. For example, even small spatial offsets in forecasted convection could locate the convection outside the radar range, normally verifying as a miss instead of a hit.

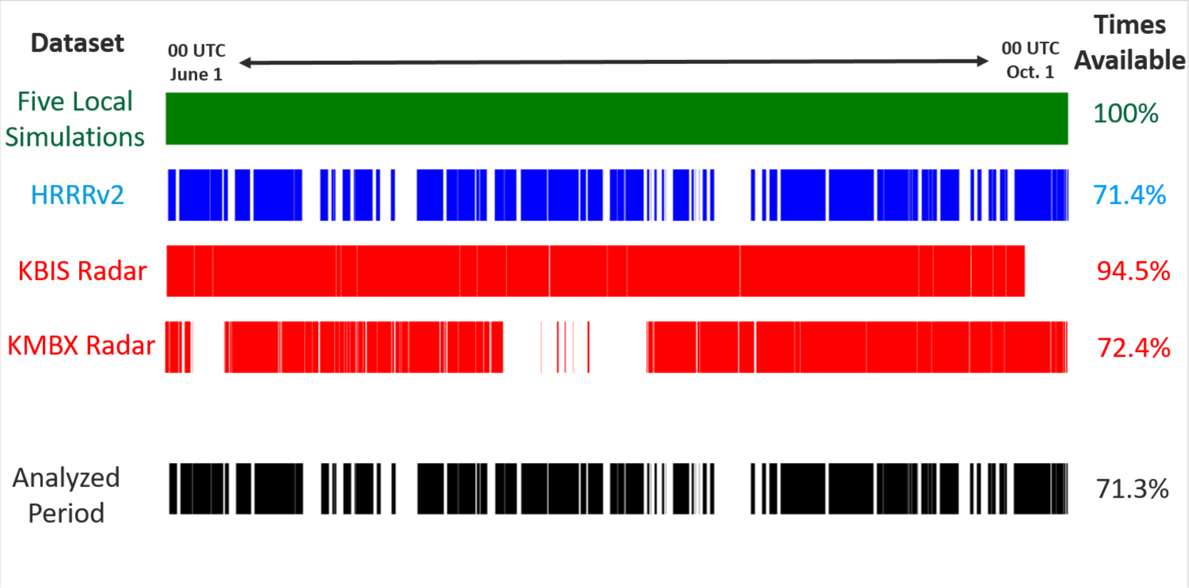

Figure 2. Hourly data availability across the model evaluation period.

To enable a fair comparison, the forecasts and observations are constrained temporally and spatially to ensure that both are sampling the same potential phenomena. While the local WRF runs have complete data coverage over the forecast period, the HRRR forecasts and radar observations contain several periods of missing data (Fig. 2). Therefore, for forecasts to be evaluated, all three datasets must be available at each hourly comparison. When data are present for all forecasts and either radar, the domain of the forecasts is masked to only include the data within the 126 km range of the radar(s). All forecast data outside the 126 km radar range are excluded from the analysis in order to ensure spatial consistency. For cases where observations are only available from one radar, the analysis is still performed but the forecasts are constrained to the range of that single radar.

The evaluation is performed by utilizing concepts from object-based verification procedures (e.g., Davis et al. 2006a). The object-based approach defines "objects" as discrete regions of interest, as in Davis et al. 2006a. For this study, objects are defined as discrete convective cells generated by applying different intensity (i.e., reflectivity) thresholds to the simulated and radar reflectivity fields. Object counts and attributes such as area and reflectivity distribution within objects are retained for analysis.

a) Bulk Analysis of Convective Objects

A time series of the bulk sum of objects across the entire forecast period at the 2-km height is presented in Fig. 3. At the 5 dBZ threshold (Fig. 3a), the Thompson forecasts match observations relatively well. WSM6 forecasts significantly over-predicted the amount of weaker convection across the entire period, with the most pronounced bias during the diurnal cycle. All forecasts overly intensify diurnal cycle convection, with significant over-prediction of objects at the 30 and 45 dBZ thresholds (Fig. 3b,c). The local WRF forecasts also predict the peak in diurnally-driven convection earlier than observed by one hour, and the HRRR peak is two hours earlier (Fig. 3c). All forecasts also cease strong convective activity too early, even if the offset in diurnal convection is accounted for (Fig. 3c). Practically all Thompson forecasts (regardless of intensity threshold) generate less objects than WSM6 forecasts; however, the differences are less pronounced at higher intensities (i.e., 30 and 45 dBZ) where no model setup is considerably outperforming the others. Outside the diurnal cycle, Thompson forecasts actually under-predict object counts at 30 dBZ (Fig. 3b).

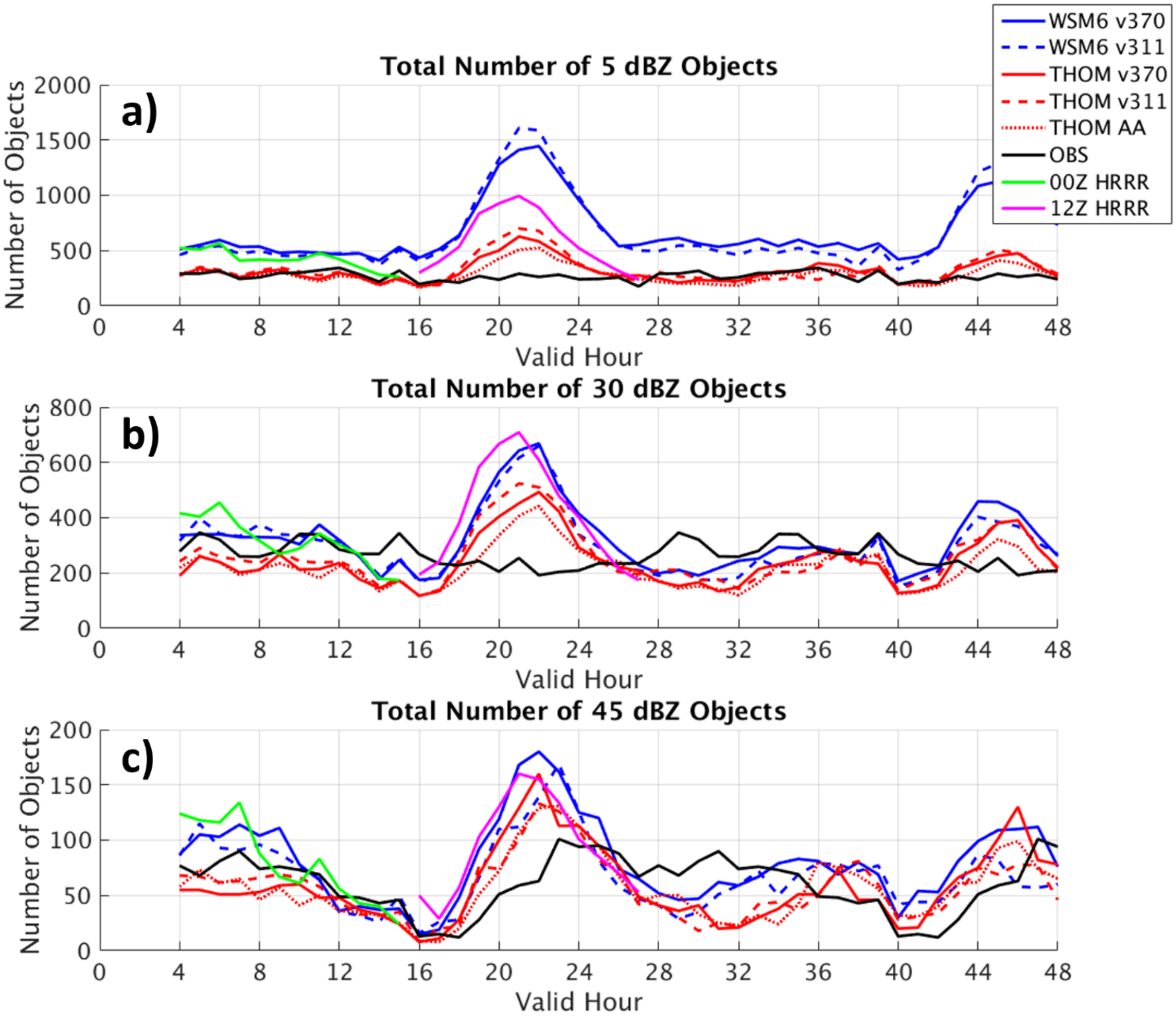

Figure 3. A time-series of the total number of a) 5, b) 30, and c) 45 dBZ objects for (blue) WSM6, (red) Thompson, (green) 00 UTC HRRR, (magenta) 12 UTC HRRR, and (black) radar observations across the entire evaluation period. Dashed lines indicate different versions of microphysics.

A time series of the total area covered by all objects at the 2-km height (similar to Fig. 3) is shown in Fig. 4. For all object reflectivity thresholds, WSM6 objects cover significantly more area than the other forecasts and covers at least double the area of observations except in the late evening and early morning hours (i.e., valid hours 28-32; 11 pm-3 am local time). HRRR and Thompson forecasts have a much better representation of areal coverage, with the Thompson forecasts being very close to observations for the first 24 hours at 45 dBZ (Fig. 4c). There is a significant difference in area between forecasts at the start of the forecast period (i.e., 04 to 09 UTC), which is nearly not present 24-hrs later. This difference may be influenced by either model initialization or skill in predicting overnight MCSs. It is worth noting, that model initialization and associated spin-up occurs at 00 UTC, which is commonly when storms are present and grow upscale or MCSs are already present in the region. A subjective investigation of several MCSs showed that when this was the case, the simulated

MCS lagged behind the observed MCS by approximately two hours, as the model attempted to correctly develop internal convective dynamics before the MCS could propagate forward.

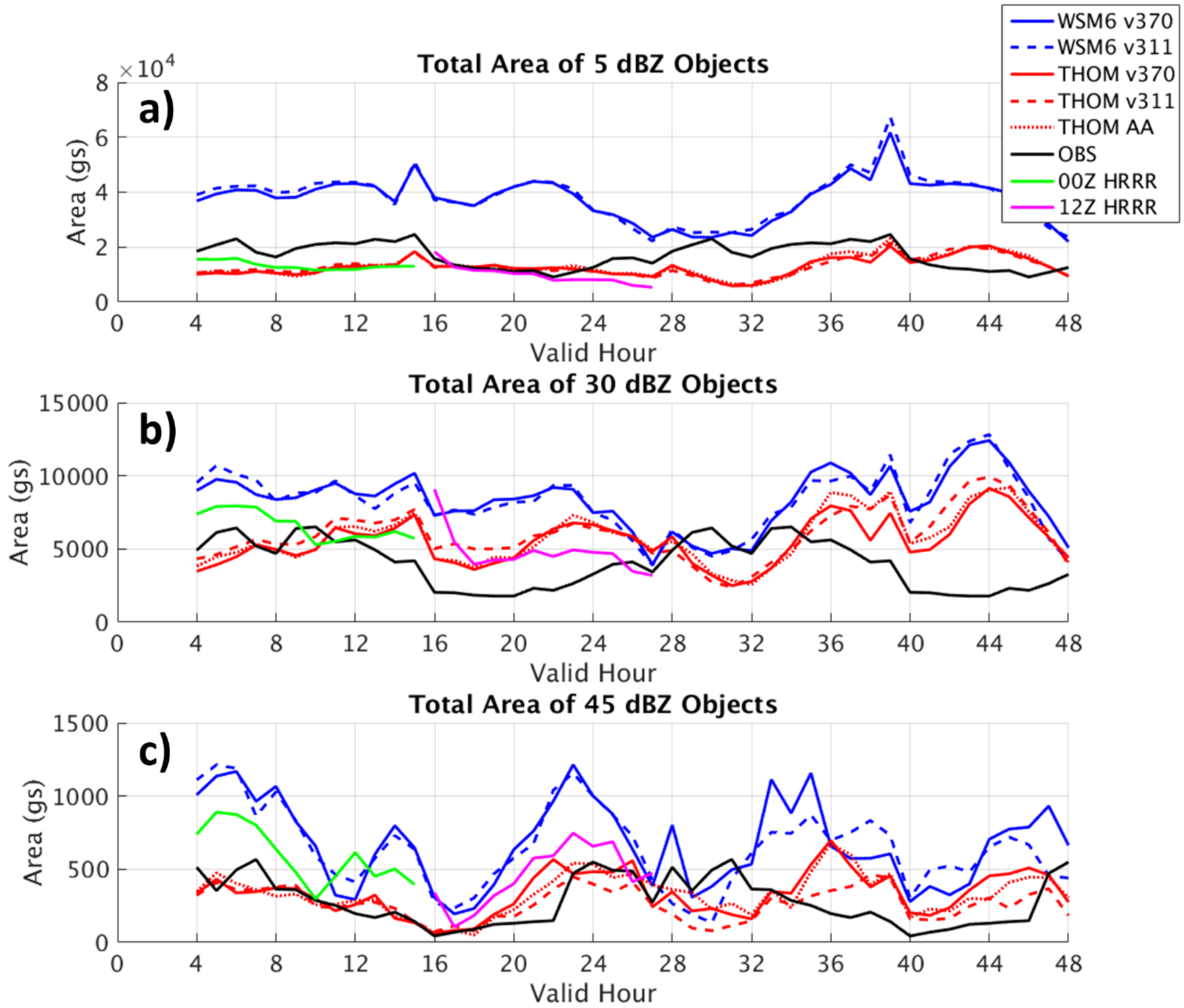

Figure 4. As in Fig. 3, except total sum of object area across the entire evaluation period.

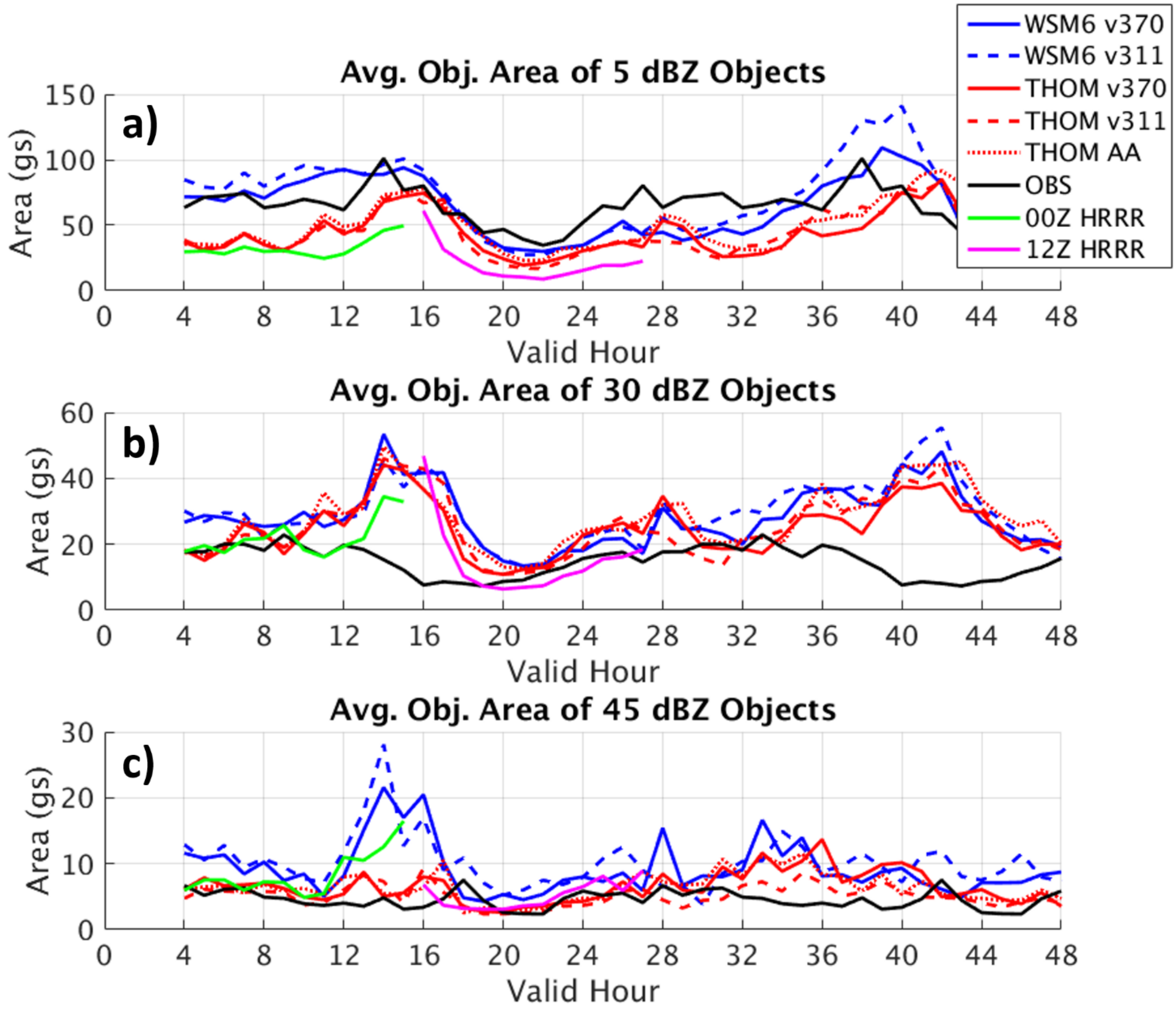

Figure 5. As in Fig. 3, except for the average object area determined using bulk object number and bulk object area values.

By analyzing the average object area (Fig. 5), it is apparent that the large total area of objects generated by the WSM6 scheme (Fig. 4) is caused by the scheme generating too many objects (especially visible at 5 dBZ). In particular, at 5 dBZ WSM6 has the average object area that matches observations the closest (Fig. 5a). All model configurations produce very similarly sized objects across the forecast period at 30 and 45 dBZ (Fig. 5b,c). Furthermore, all forecasts contain a large spike in average object areas between 12 and 18 UTC at 30 dBZ and only the Thompson scheme does not contain such a spike at the 45 dBZ threshold. Forecasts are predicting objects between two and four times larger than observations during this time.

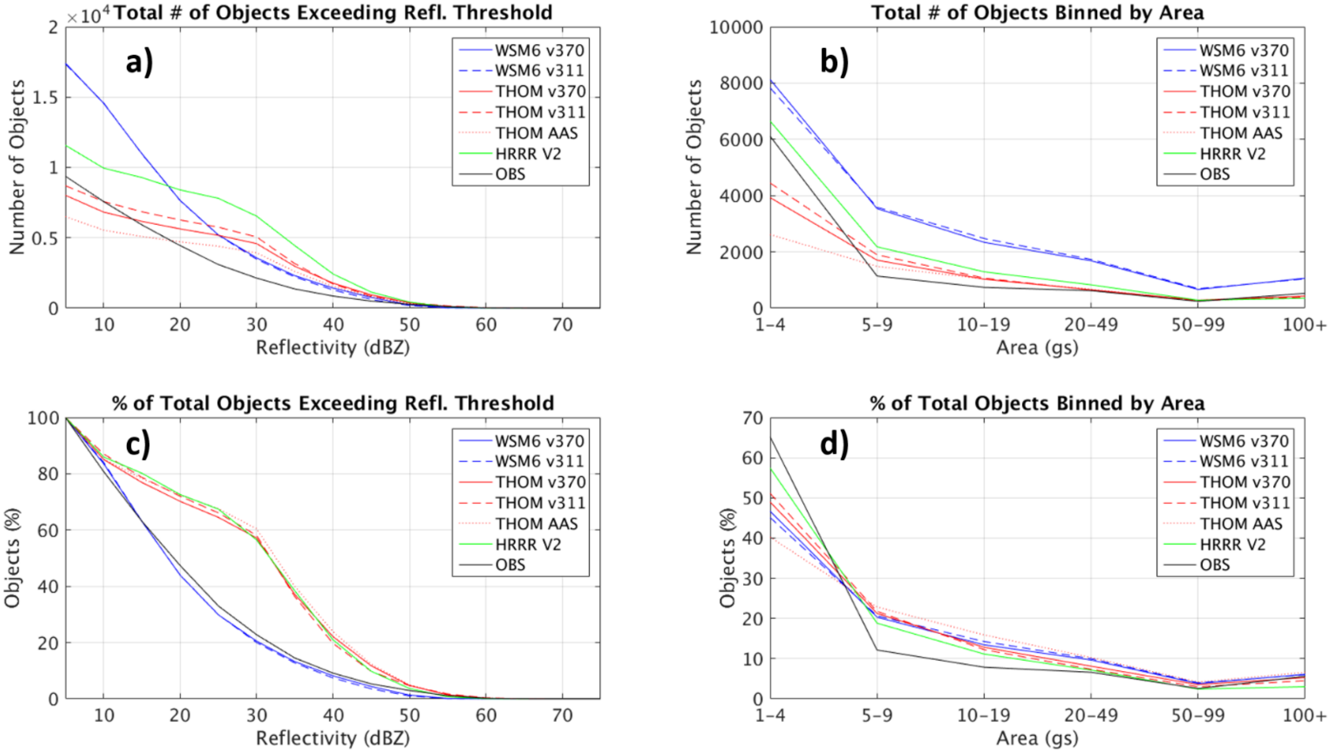

Figure 6. The total number of 5 dBZ objects a) with reflectivity values exceeding a certain threshold and b) binned by area; the percentage of 5 dBZ objects c) with reflectivity values exceeding a certain threshold and d) in each area bin at the 2-km height.

While WSM6 contains the largest number of 5 dBZ objects and significantly more than observations (Fig. 6a), the percentage of objects that exceed different reflectivity thresholds is very similar to observations (Fig. 6c). For example, only 20% of 5 dBZ WSM6 objects contain reflectivity values that exceed 30 dBZ, while 60% of 5 dBZ Thompson and HRRR objects contain values above 30 dBZ (as compared to 23% in observations). Therefore, while WSM6 tends to generate more objects, HRRR and Thompson objects tend to be stronger and more commonly reach between 20 to 40 dBZ. Binning all 5 dBZ objects by area, objects of all area sizes are over-forecasted except the Thompson forecasts in the smallest area bin (Fig. 6b). When normalized by the total number of objects, observations only contain more objects in the smallest, 1-4 grid square (i.e., 9 to 36 km^2) bin. Forecasts contain too many large objects (45 to 891 km^2), but forecasts and observations begin to converge for objects with areas of 891 km^2 and greater due to lower frequency of such events (and the smaller spatial domain).

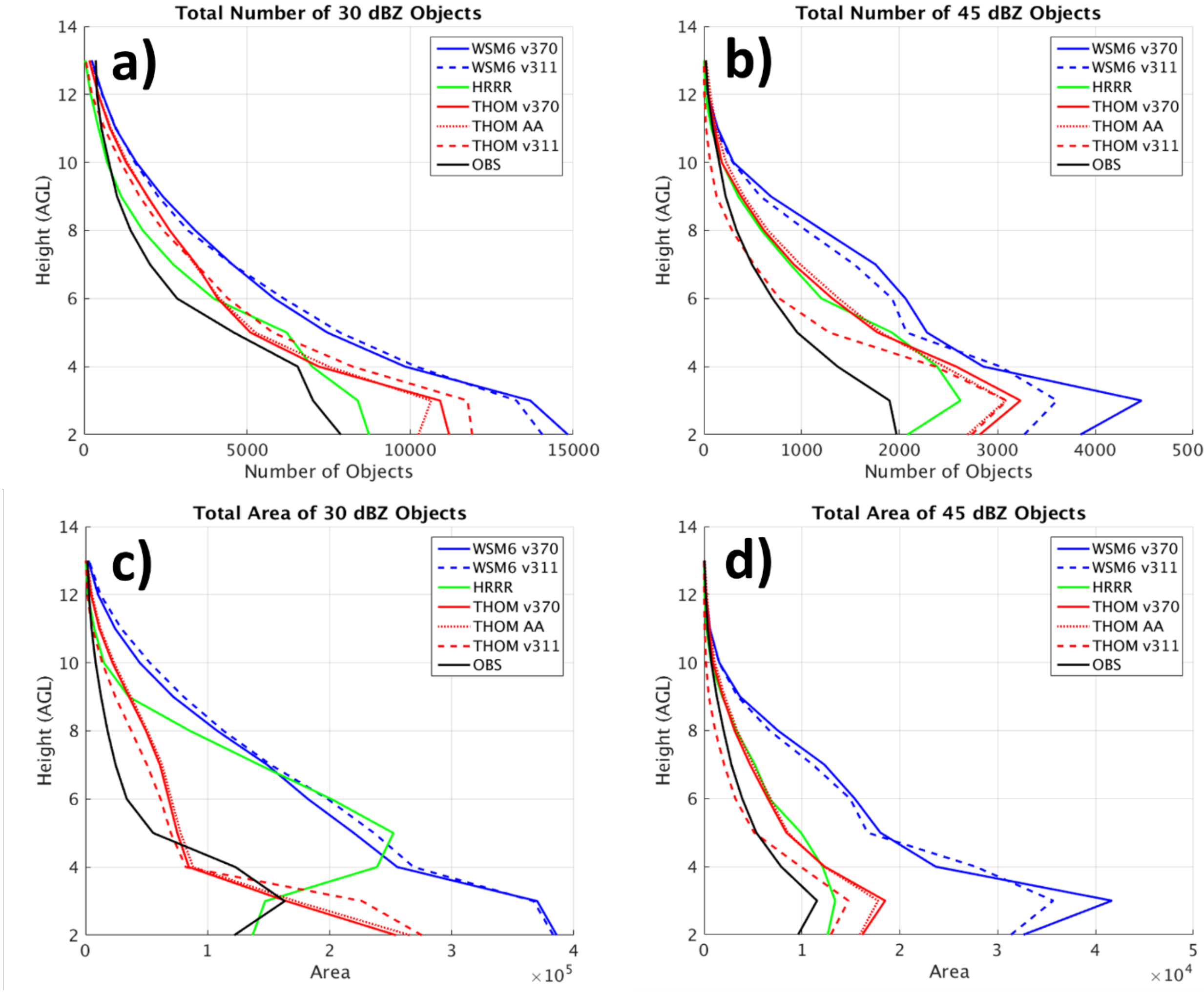

Figure 7. The total a) number and c) area of 30 dBZ objects, and the total b) number and d) area of 45 dBZ objects between 2 and 13 km AGL height.

The vertical distribution of objects and their areas in potential convection regions across the peak in convective activity (i.e., 17 to 27 UTC) is shown in Fig. 7 for the 30 and 45 dBZ thresholds. For both thresholds, all forecasts generate too many objects at all heights, except the Thompson v3.1.1 forecasts above 5 km in height for the 45 dBZ threshold (Fig. 7a,b) and HRRR forecasts above 8 km for the 30 dBZ threshold. There is a significant over-prediction of objects near the surface, with the largest over-prediction occurring at the melting level. The melting level is generally found at 3 km AGL, and is visible as a “spike” in objects and area (Fig. 7b and 7d) due to enhanced reflectivity associated with melting snow hydrometeors. The regions of enhanced reflectivity more frequently surpass the 30 and 45 dBZ thresholds, causing the increased number and area of objects. The Thompson forecasts perform very well in predicting object areas, closely matching observations at both the 30 and 45 dBZ thresholds (Fig. 7c,d). WSM6 forecasts perform considerably worse, with the largest over-prediction of object areas occurring at the melting level and corresponding with the over-prediction of objects as in Fig. 5. Interestingly, the HRRR forecasts have a significant peak in 30 dBZ object area at 4 to 6 km AGL, which is not visible in observations and is above the melting level (Fig. 7c). This peak is also not pronounced in object counts (Fig. 7a), which implies that ≥ 30 dBZ objects are significantly larger in that layer. The peak is not present at the 45 dBZ threshold (Fig. 7d), likely indicating that the peak is not directly caused by convection motions but likely involved in stratiform rain-producing processes.

b) Evaluation of Convective Processes

To evaluate the convective processes occurring in simulated and observed convection, the reflectivity magnitude distributions within objects are analyzed. Contoured frequency by altitude diagrams (CFADs) are generated for 30 and 45 dBZ objects created at 2 km. The area covered by the object at 2 km is extended up through 13 km and the reflectivity distribution at each height within the object area is analyzed. This method ensures we are only looking within objects that have precipitation reaching the ground and avoiding stringent thresholding aloft. CFADs of reflectivity values within objects ≥ 30 dBZ generated at the 2-km height are shown in Fig. 8. The differences between the model CFADs and radar CFAD (i.e., model frequency minus radar frequency) indicate that forecasts generally contain a wider spread of reflectivity values than are observed by radar (red shading; Fig. 8e,g,f). The Thompson forecasts contain a weak bias around the height of the melting level (i.e., 3 km), but contain a strong bias above the melting level with reflectivity differences of 10 to 15 dBZ occurring 10-20% of the time above 7 km (Fig. 8f). The HRRR forecasts contain a pronounced increase in the frequency of higher reflectivity values up to 5 km, which is not visible in other forecasts or observations (Fig. 8g). The initial increase in intensity leads to an offset between the HRRR CFADs and radar CFAD (Fig. 8g and 8d, respectively), where the HRRR forecasts are overly intense aloft.

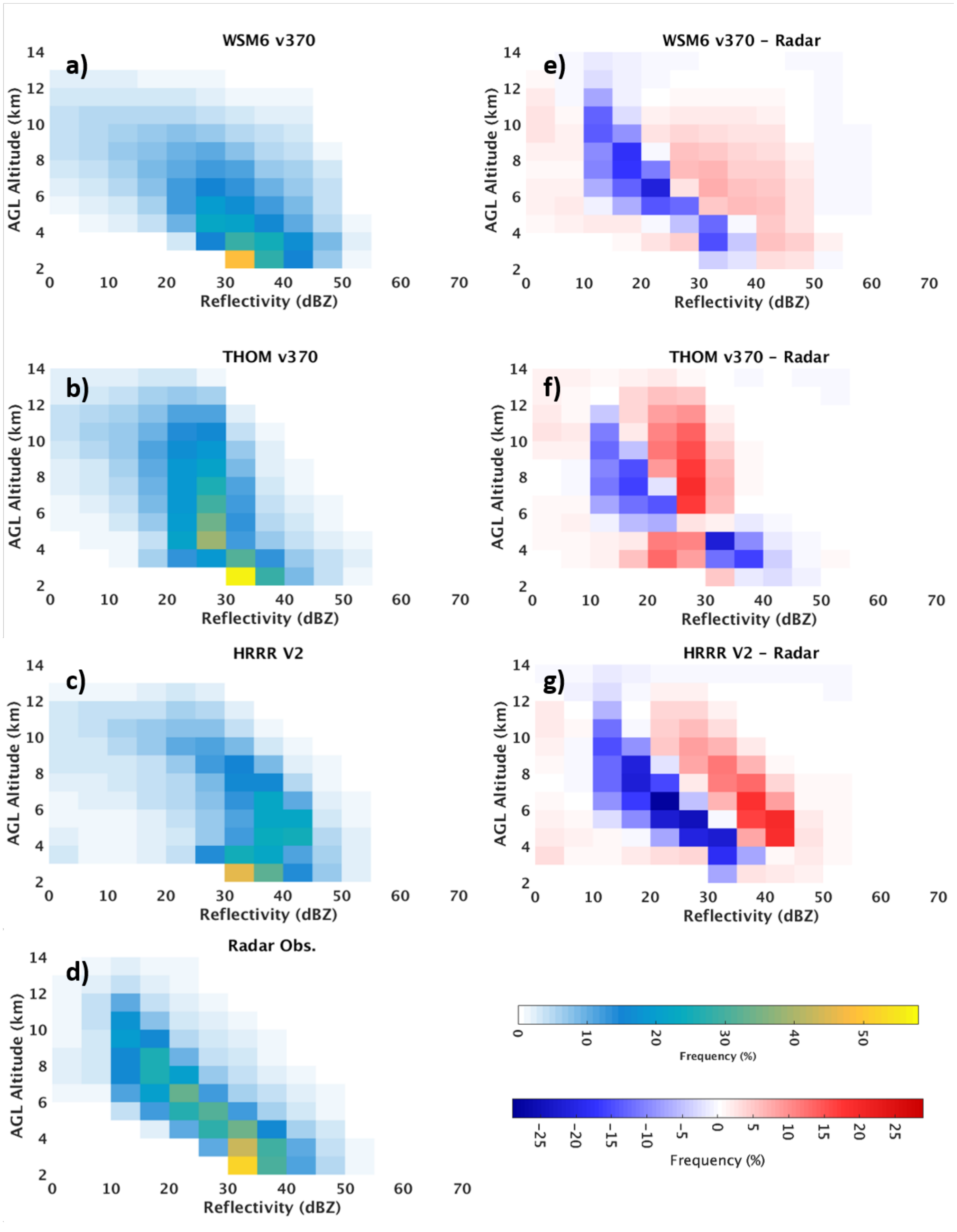

Figure 8. Contoured frequency by altitude diagrams of reflectivity, binned by every 5 dBZ, within bounds of 30 dBZ objects located at the 2-km height for a) WSM6, b) Thompson, c) HRRR, and d) radar observations. Differences between the simulated CFADs and radar CFAD (i.e., forecast minus observation) are presented in e-g, where red indicates a positive bias (over-prediction) and blue indicates a negative bias (under-prediction).

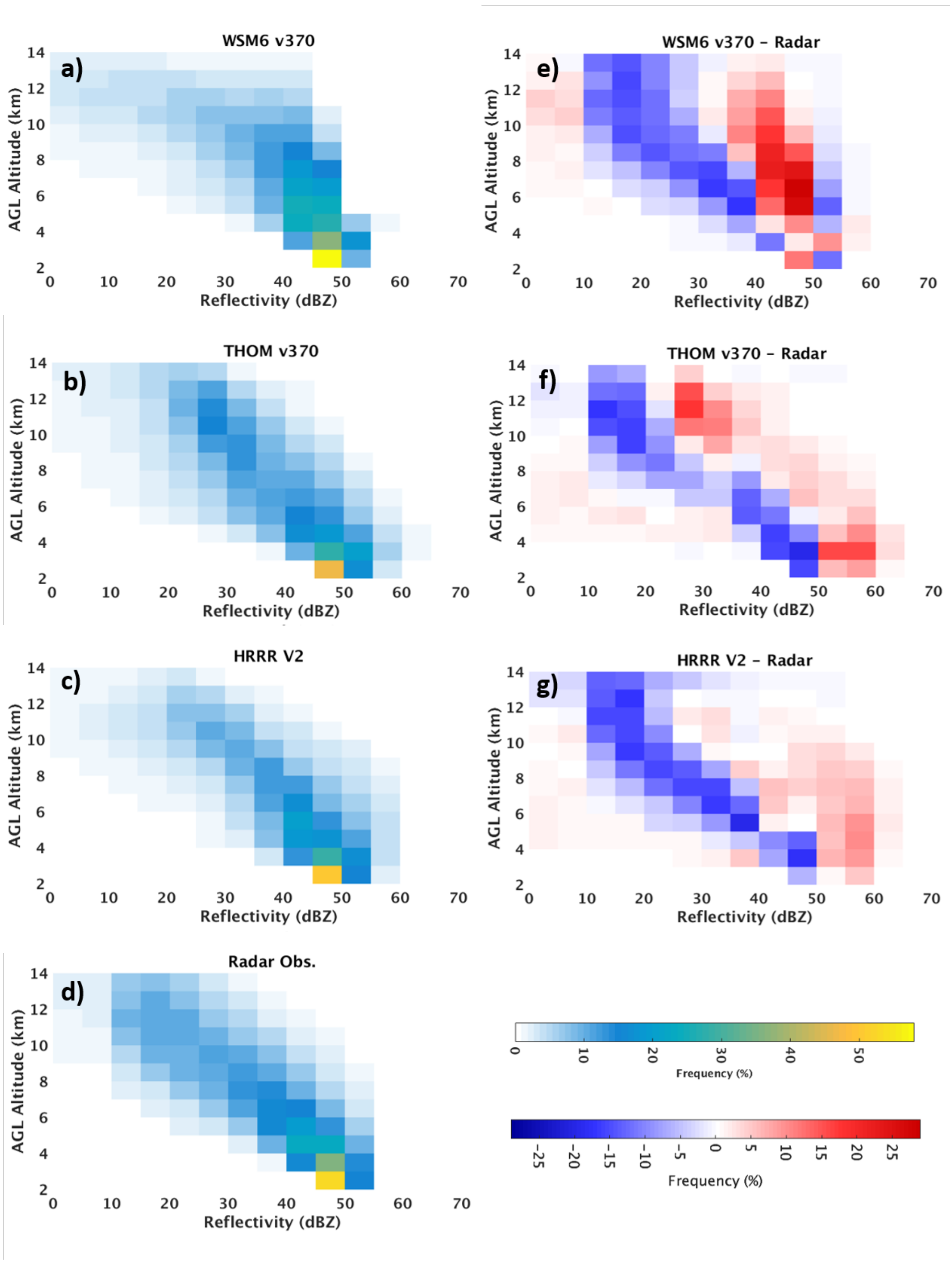

Figure 9. As in Fig. 8, except the reflectivity distribution is taken with 45 dBZ objects at 2 km.

The ≥ 45 dBZ object model CFADs contain less spread and are more consistent with each other and observations (Fig. 9a-d) than the 30 dBZ object CFADs. The slope of the WSM6 CFAD decreases too slowly between the melting level and 10 km when compared to observations, indicating convection that is generally too strong (Fig. 9e). The slope of the Thompson frequency distribution matches the radar observations well, but is overly intense again by 10 to 15 dBZ both near the melting layer and aloft (~10-13 km; Fig. 9f). HRRR forecasts have a strong bias up to 9 km, although not as pronounced as the Thompson forecasts (Fig. 9g); however, the HRRR under-predicts the frequency of the deepest storms (11-14 km), suggesting HRRR under-predicts storm depth for a number of events.

c) Model Intercomparison

Model realizations with varying versions are compared to determine how model forecasts have changed in time for the purpose of operational significance. Differences in the Thompson CFADs are presented in Fig. 10. Comparing the Thompson v3.1.1 forecasts against the Thompson 3.7.0 forecasts, it is apparent that for the 5 and 30 dBZ object thresholds, the most pronounced differences occur at the height of the melting level (Fig. 10a,b). Thompson v3.7.0 frequently contains weaker reflectivity values around the melting level, indicating a more robust separation between stratiform and convective precipitation. The Thompson v3.7.0 forecasts are more intense than in v3.1.1 in deep convection, with a much higher frequency of higher reflectivity values above the melting level of up to 15% (Fig. 10c).

The differences between the TAA v3.7.0 and Thomson v.3.7.0 forecasts are less significant than the differences between the Thompson v3.1.1 and v3.7.0 forecasts. At 5 dBZ and 30 dBZ, the maximum differences in frequency of up to 2% are negligible (Fig. 10d). The largest differences are found at and below the melting level for objects ≥ 45 dBZ. The TAA scheme is less intense near the surface, with differences up to 15 dBZ present 4% of the time (Fig. 10f). The TAA scheme contains higher reflectivity values aloft; however, the differences in frequency are even lower (< 2%).

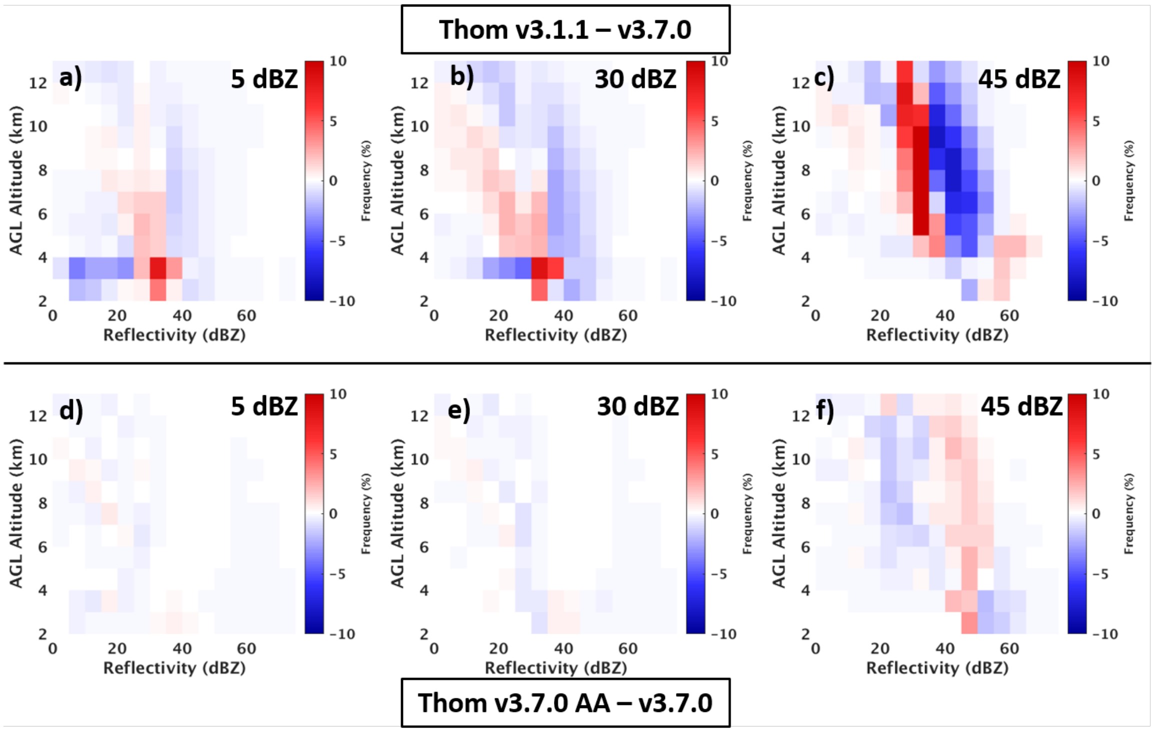

Figure 10. Differences in CFADs between (top) Thompson v3.1.1 and v3.7.0 and (bottom) Thompson aerosol-aware v3.7.0 and v3.7.0. For Thompson (top), blue indicates regions where v3.7.0 contains a higher frequency than v3.1.1. For Thompson aerosol-aware (bottom), blue indicates regions where v3.7.0 contains a higher frequency than the aerosol-aware scheme.

The comparison between WSM6 versions are presented in Fig. 11. There are very minor differences near the surface (i.e., 2 km) at all 5, 30, and 45 dBZ object thresholds, which echoes the conclusion reached in Fig. 3; however, more significant differences are found aloft that would be missed if evaluation in the vertical was not included. It was shown that WSM6 v3.7.0 contained a strong bias aloft when compared to observations (Fig. 9e); however, WSM6 v3.1.1 is frequently more intense than v3.7.0 at those heights (Fig. 11c). Therefore, for deep convection WSM6 v3.7.0 has improved upon the strong bias in v3.1.1, leading to a decreased frequency of higher reflectivity values above the melting level (Fig. 11c).

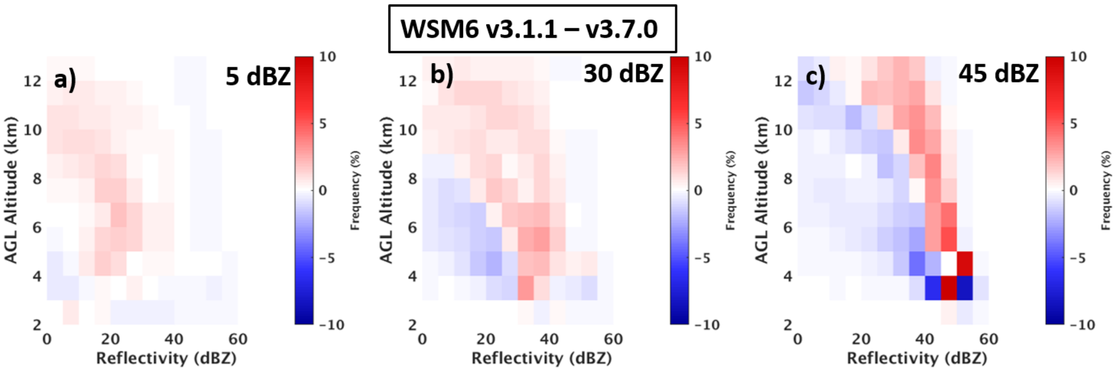

Figure 11. As in Fig. 9, except for the frequency difference between WSM6 v3.1.1 and v3.7.0 forecasts. Blue shading indicates higher frequency of occurrence in v3.7.0 forecasts.

4. Summary and Discussion

A group of high resolution convection-allowing summertime forecasts were evaluated by comparing the simulated reflectivity field against the radar reflectivity field. An object-based evaluation was performed at the 2-km height to determine the number of convective objects and their respective areas at different reflectivity thresholds of ≥ 5, 30 and 45 dBZ. At 5 dBZ, WSM6 significantly over-predicted convective objects during the entire period, while the Thompson and HRRR forecasts performed reasonably well. The WSM6 scheme also over-forecasts the area covered by convection, which was due to the large number of forecasted objects. During the diurnal maximum in convective activity, all forecasts generated significantly too many objects. At 45 dBZ, there is a clear offset in the timing of convection, where the local WRF forecasts and HRRR forecasts reach the peak in convective activity one and two hours earlier than observations, respectively. The evaluation was expanded into the vertical, by performing a bulk analysis of potentially convective (i.e., 30 and 45 dBZ) objects at each height from 2 to 13 km. All forecasts over-predicted object counts at practically all heights, with the greatest biases being present near the surface and melting level. All forecasts over-predicted total area coverage with height except the Thompson forecasts, which performed very well even though total object-counts were over-predicted. A well-defined maximum in area in the HRRR forecasts that is 1 and 3 km above the typical melting level height was found, which was not seen in observations or the other forecasts and requires further investigation.

The CFADs of reflectivity for the model forecasts were compared against the radar reflectivity CFAD to determine if there are any substantial differences between convective processes occurring in forecasts and observations. At 30 dBZ, forecasts contain a much wider spread of reflectivity values than observations. The Thompson forecasts were found to have a weak bias below the melting level and significant strong bias above the melting level, which may be resultant of the reflectivity calculation and snow hydrometeors in the Thompson scheme. Further investigation into the sensitivity of simulated reflectivity calculation needs to be performed. The HRRR is the only member that had an increase in intensity with height before weakening, resulting in an offset from observations in the locations of highest frequencies. This result may be influenced by the simulated reflectivity calculation used, and requires further investigation. Model performance improved at 45 dBZ. WSM6 forecasts contained a pronounced strong bias above the melting level. Although the slope of the highest reflectivity frequencies in the HRRR and Thompson forecasts matched observations, there was an offset in intensity resulting in a strong bias.

An intercomparison between model versions was performed to identify if any operationally-significant changes have occurred. Comparing Thompson v3.1.1 against v.3.7.0, showed the convection in v3.7.0 is more intense, especially aloft (i.e., 6-10 km) in deeper convection. There was also a more distinct separation between stratiform and convective precipitation at the melting level, where v3.7.0 has a much higher frequency of lower reflectivity values. No considerable differences were found between the Thompson and Thompson TAA schemes. Comparing WSM6 v3.1.1 against v3.7.0 showed that while no notable differences exist near the surface (i.e., the 2-km level), v3.7.0 improved upon the strong bias that was present aloft in v3.1.1. An analysis focusing only on one height would miss the differences aloft; therefore, analyses that incorporate the entire vertical structure are needed. This methodology also showcases a way for evaluating processes occurring in model forecasts after version or physics changes.

The overall bulk statistics show that the model forecasts evaluated are generally overly convective, leading to an excess of convective objects and/or area. These biases are present across all heights, but are dominant near the surface and at the melting level. There were no consistent differences in skill between the local run forecasts and the operational HRRR, at least in a bulk analysis. The CFADs showed that while model forecasts are able to correctly simulate convective processes, they contain a significant spread that is not seen in observations.

References

Casati, B., G. Ross, and D. B. Stephenson, 2004: A new intensity-scale approach for the verification of spatial precipitation forecast. *Meteor. Appl.*, **11**, 141-154.

Clark, A. J., R. G. Bullock, T. L. Jensen, M. Xue, and F. Kong, 2014: Application of object-based time-domain diagnostics for tracking precipitation systems in convection-allowing models. *Wea. Forecasting*, **29**, 517-542.

Davis, C., B. Brown, and R. Bullock, 2006a: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev.*, **134**, 1772-1784.

Davis, C., B. Brown, and R. Bullock, 2006b: Object-based verification of precipitation forecasts. Part II: Application to convective rain systems. *Mon. Wea. Rev.*, **134**, 1785-1795.

Ebert, E. E., and J. L. McBride, 2000: Verification of precipitation in weather systems: Determination of systematic errors. *J. Hydrology*, **239**, 179-202.

Koch, S. E., B. S. Ferrier, M. T. Stoelinga, E. J., Szoke, S. J. Weiss, and J. S. Kain, 2005: The use of simulated radar reflectivity fields in the diagnosis of mesoscale phenomena for high-resolution WRF model forecasts. *11th Conf. On Mesoscale Processes,* Albuquerque, NM, Amer. Meteor. Soc., J4J.7, [Available online at http://ams.confex.com/ams/pdfpapers/97032.pdf].

Mass, C. F., D. Ovens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bull. Amer. Meteor. Soc.*, **83**, 407-430.

Tang, L., J. Zhang, C. Langston, J. Krause, K. Howard, and V. Lakshmanan, 2014: A Physically Based Precipitation–Nonprecipitation Radar Echo Classifier Using Polarimetric and Environmental Data in a Real-Time National System. *Wea. Forecasting,* **29**, 1106–1119, https://doi.org/10.1175/WAF-D-13-00072.1.

Thompson, G. and T. Eidhammer, 2014: A Study of Aerosol Impacts on Clouds and Precipitation Development in a Large Winter Cyclone. *J. Atmos. Sci.,* **71**, 3636–3658, doi: 10.1175/JAS-D-13-0305.1.