Verification and Model Configuration Sensitivity of Simulated ABI Radiance Forecasts With the FV3-LAM Model

Aaron Johnson[1] [iD], Xuguang Wang[1] [iD], Benjamin T. Blake[2,3], Eric Rogers[3], Yunheng Wang[4,5], Jacob R. Carley[3] [iD], Adam Clark[4], Jeffrey Beck[6,7], and Curtis Alexander[6]

[1]School of Meteorology, University of Oklahoma, Norman, OK, USA, [2]SAIC, Reston, VA, USA, [3]NOAA National Centers for Environmental Prediction, Environmental Modeling Center, College Park, MD, USA, [4]NOAA National Severe Storms Laboratory, Norman, OK, USA, [5]Cooperative Institute for Severe and High-Impact Weather Research and Operations, University of Oklahoma, Norman, OK, USA, [6]NOAA Global Systems Laboratory, Boulder, CO, USA, [7]Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, CO, USA

Abstract This study evaluates simulated radiance forecasts from a series of controlled experiments consisting of FV3-LAM forecasts with different configurations of model physics and vertical resolution. The forecasts were produced during the 2020 Hazardous Weather Testbed Spring Forecasting Experiments on the same forecast cases. The evaluation includes grid-point, neighborhood-based and object-based verification. The experiments include forecasts that were identical except for the physics (EMC-LAM vs. EMC-LAMx), vertical resolution (EMC-LAMx vs. NSSL-LAM), or combined initial conditions, physics and vertical resolution (GSL-LAM). It is found that the EMC-LAM generally provided better simulated radiance forecasts than the other three configurations at most forecast lead times, due to its unique physics configuration. All configurations generally over-forecasted high level clouds. EMC-LAM reduced the over-forecasting of high clouds, but also under-forecasted the coverage of mid-level clouds. In contrast, at early lead times the EMC-LAM had relatively poor performance relative to the other forecasts. Furthermore, EMC-LAM was an outlier in terms of the vertical structure of clouds. It is also found that the NSSL-LAM consistently improved upon the EMC-LAMx, which had fewer vertical levels than NSSL-LAM. Compared to EMC-LAMx, NSSL-LAM had less cloud over-forecasting bias, especially with small cloud objects, and less overall error. The differences between EMC-LAMx and GSL-LAM were generally much smaller than the differences between EMC-LAMx and EMC-LAM/NSSL-LAM. Finally, it is found that a non-linear bias correction conditioned on symmetric brightness temperature reduced the overall root-mean-square error by about a factor of 2 while improving the unrealistic vertical structure of clouds in the EMC-LAM.

Plain Language Summary While forecasts from different convection-allowing model (CAM) configurations are often evaluated in terms of simulated reflectivity or precipitation, many users are also interested in simulated brightness temperature and associated cloud features. This study evaluates the forecast performance of several configurations of the FV3-LAM CAM in terms of simulated infrared brightness temperature. The different configurations highlight the impacts of vertical resolution and physical parameterizations, in particular. It is found that the older physics parameterization, with Geophysical Fluid Dynamics Laboratory microphysics and Eddy-Diffusivity Mass-Flux boundary layer, performed generally better than a newer configuration, with Thompson microphysics and Mellor-Yamada-Nakanishi-Niino boundary layer, that had previously been found to improve precipitation and simulated reflectivity forecasts. The difference is related to both the presence of an additional cumulus parameterization in the older physics, as well as the tendency of the newer Thompson microphysics scheme to produce large regions of high cloud near deep convection. However, at early lead times the older physics configuration took longer to spin up ongoing convection. It is also found that enhanced vertical resolution consistently improved the forecasts. A caveat to these conclusions is that the FV3-LAM and its physics configuration are undergoing continuous development which may change these results in the newest model versions.

1. Introduction

The National Centers for Environmental Prediction (NCEP) is currently developing the Rapid Refresh Forecast System (RRFS; Carley et al., 2021) as part of an effort to unify the NCEP operational forecast modeling systems using the finite volume cubed sphere (FV3; Putman & Lin, 2007) dynamical core. The RRFS is planned to run

at a convection-allowing grid spacing of 3 km using the FV3 limited area model (LAM; Black et al., 2021). The RRFS will replace current NCEP operational convection-allowing model (CAM) systems that are based on the Weather Research and Forecasting (WRF; Skamarock et al., 2008) and Nonhydrostatic Multiscale Model on the B-grid (Janjić & Gall, 2012; Janjić et al., 2001) dynamic cores which have been extensively evaluated, developed and tuned as a CAM over the years (e.g., Aligo et al., 2018; Clark et al., 2015; Grasso et al., 2014; Johnson & Wang, 2017, 2019; Johnson, Wang, Xue, & Kong, 2011; Kain et al., 2008; Weisman et al., 2008).

Several studies have begun to evaluate different configurations of the FV3-LAM as a CAM in order to facilitate continued development of the FV3-LAM model and physics parameterizations. For example, Gallo et al. (2021) applied neighborhood and object-based methods to evaluate FV3-based CAM forecasts of updraft helicity and reflectivity. They found that the FV3-based CAM forecasts with several configurations were generally less skillful than a WRF-based high resolution rapid refresh (HRRR; Dowell et al., 2022; James et al., 2022) model. Potvin et al. (2019) also found two FV3 configurations to underperform a variety of WRF configurations for severe storm surrogate forecasts, although FV3 did have some advantages over WRF in terms of storm structure. However, Zhou et al. (2019) found an improved precipitation diurnal cycle in FV3-based CAM forecasts, compared to FV3 forecasts at 13 km resolution, while Snook et al. (2019) found improved quantitative precipitation forecasts in an FV3-based CAM configuration, compared to several configurations of WRF-based CAM forecasts. The FV3 and WRF forecasts in these studies generally used different physics and initial conditions in addition to the different dynamical cores.

An aspect of the forecasts with FV3-based CAM, and its associated physical parameterizations, that has received relatively little attention in the literature is the forecast of simulated satellite radiance. Evaluation of simulated all-sky radiance can provide information on how well the model simulates cloud processes beyond precipitating deep convection. The performance of forecast cloud processes is important to understand because clouds affect the radiation budget (e.g., surface heating) as well as latent heating of condensation during cloud formation, both of which can impact forecasts on time scales of a day or less. Furthermore, studies have shown the potential to improve CAM forecasts through assimilation of the high time and space resolution radiances from the Advanced Baseline Imager (ABI; Schmit et al., 2017) onboard the geostationary GOES-16 and GOES-17 satellites (e.g., Johnson et al., 2022; Jones et al., 2020; Zhang et al., 2021). The fidelity of model-simulated and observed all-sky ABI radiances should be considered for optimal assimilation of such radiances and should therefore also be evaluated.

An evaluation of simulated all-sky radiances from different WRF configurations found that the microphysics parameterization scheme had more pronounced impacts on the simulated radiance forecasts than planetary boundary layer (PBL) parameterization (Cintineo et al., 2014), especially for upper-level clouds (Otkin & Greenwald, 2008). In addition to grid-point and neighborhood-based diagnostics, object-based evaluation has also proven effective for quantifying physical differences between model-simulated and observed cloud features (Griffin et al., 2017; Rempel et al., 2017). Similar evaluations of the interactions between simulated cloud characteristics and physics configuration have not been performed for FV3-based CAM forecasts. Since FV3 has a fundamentally different dynamical core and nonhydrostatic solver (Harris et al., 2020), and different available physics options than WRF, the first goal of the present study is to evaluate the simulated ABI all-sky radiances in a series of FV3-based CAM forecasts with different physics configurations using both grid-point and object-based diagnostics. In particular, we aim to better understand how well current FV3 physics schemes forecast simulated radiance, and to better understand the relative impact of the FV3 physics configuration on simulated radiance forecasts relative to other aspects of the model configuration such as initial condition (IC) source.

The model vertical resolution may also influence the simulated radiance forecasts. This aspect of the model configuration in CAM forecasts has received less attention in the literature than the physics configuration and IC. However, the vertical resolution may be an important consideration in the context of simulated radiance evaluation and data assimilation because the vertical gradients present at cloud top where the observed radiance weighting function peaks may not be fully resolved at typical vertical resolutions on the order of 500 m in the middle and upper troposphere (e.g., Johnson & Wang, 2019). Furthermore, the vertical resolution may influence the evolution of deep convection systems that often dominate the cloudy radiances in the warm season through interactions with the microphysics scheme (e.g., Aligo et al., 2009). Therefore, we also evaluate the forecasts with different vertical resolutions with a goal of better understanding the impacts of vertical resolution on the FV3-simulated ABI radiances, particularly in cloudy regions.

Table 1
List of Forecast Cases Used in This Study

1 May 2020	11 May 2020
3 May 2020	14 May 2020
4 May 2020	16 May 2020
5 May 2020	26 May 2020
6 May 2020	27 May 2020
8 May 2020	28 May 2020
9 May 2020	29 May 2020

The present study aims to guide future improvements to operational FV3-LAM configurations by providing a better understanding of the error characteristics of FV3-LAM simulated ABI radiances, and their sensitivity to model configuration. Such understanding can also guide ABI radiance data assimilation efforts, including the application of appropriate bias-correction techniques. Therefore, the simulated ABI radiances in the present study are also evaluated after applying a non-linear bias correction following Chandramouli et al. (2022).

In summary, this study evaluates a series of FV3-LAM forecasts with different physics and vertical resolutions in terms of simulated ABI radiance. The goal of the evaluation is to (a) quantify the performance of the simulated ABI radiance forecasts using grid-point, neighborhood-based and object-based diagnostics, (b) determine the sensitivity of simulated radiance forecasts to the above aspects of the model configuration, and (c) explore non-linear bias correction of the FV3-LAM simulated radiances and its impact on the forecast performance. The remainder of this paper is organized as follows. Section 2 describes the methods used in this study, including the experiment design and verification metrics. Results are presented in Section 3 and a summary and conclusions are provided in Section 4.

2. Methods

This study leverages forecasts run during the 2020 Hazardous Weather Testbed (HWT) Spring Forecasting Experiments by the National Oceanographic and Atmospheric Administration (NOAA) National Severe Storms Laboratory (NSSL), NOAA Environmental Modeling Center (EMC), and NOAA Global Systems Laboratory (GSL). A total of 14 forecasts, initialized at 0000 UTC, were available on the same cases for direct comparison of the impact of the different FV3-LAM configurations used for each forecast (Table 1).

2.1. Experiment Design

The FV3-LAM configurations are listed in Table 2. All forecasts used 3-km grid spacing and rapid radiative transfer model (Clough et al., 2005) radiation parameterization scheme. Forecasts labeled NSSL-LAM, EMC-LAM, and EMC-LAMx all were initialized from the NCEP Global Forecast System (GFS) analysis and used the corresponding GFS forecast as lateral boundary conditions (LBC). NSSL-LAM and EMC-LAMx both used the Thompson et al. (2008) microphysics scheme, Mellor-Yamada-Nakanishi-Niino (MYNN; Nakanishi & Niino, 2009) PBL scheme, and Noah (Barlage et al., 2010) land surface model. The only difference between NSSL-LAM and EMC-LAMx was the number of vertical levels (Table 2), providing a controlled experiment on the impact of the finer vertical resolution in NSSL-LAM. The only differences between EMC-LAMx and EMC-LAM are that EMC-LAM uses the Geophysical Fluid Dynamics Laboratory (GFDL; Chen & Lin, 2013; Zhou et al., 2019) microphysics scheme, and Eddy-Diffusivity Mass-Flux (J. Han et al., 2016) PBL scheme. This difference provides a controlled experiment for the impact of the more recent physics package used in EMC-LAMx. In addition to the GFDL microphysics scheme, the EMC-LAM configuration also includes deep cumulus parameterization using the Simplified Arakawa-Schubert (J. Han & Pan, 2011) scheme. However, only about 6.5% of the total precipitation on these cases was produced by the cumulus parameterization; most of the

Table 2
Summary of FV3-LAM Configurations, Including the Name of the Configuration Used in This Study

	IC	LBC	MP	CU	PBL	LSM	NZ
NSSL-LAM	GFS	GFSf	Thompson	None	MYNN	Noah	82
EMC-LAM	GFS	GFSf	GFDL	SAS	EDMF	Noah	60
EMC-LAMx	GFS	GFSf	Thompson	None	MYNN	Noah	60
GSL-LAM	HRRRv4	RAPf	Thompson	None	MYNN	RUC	64

Note. Source of Initial (IC) and Lateral Boundary Conditions (LBC), Microphysics Parameterization Scheme (MP), Cumulus Parameterization (CU), Planetary Boundary Layer Parameterization (PBL), Land Surface Model (LSM), and Number of Vertical Levels (NZ). Other acronyms in the table are defined in Section 2.1.
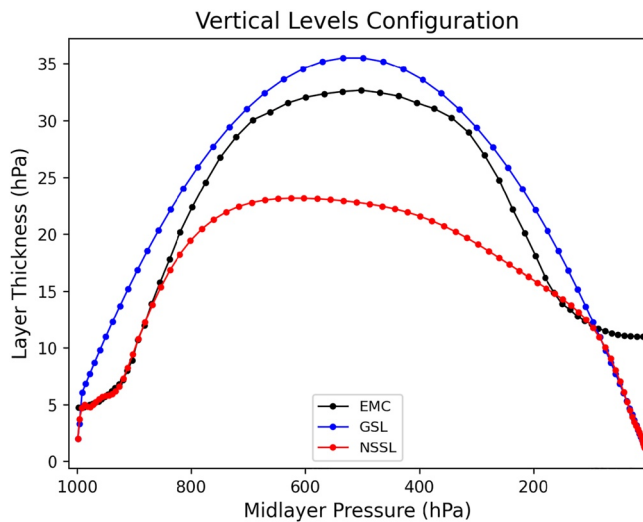
Figure 1. Differences among models in terms of the distribution of layer thickness between each vertical level. In this figure, GSL refers to the GSL-LAM, EMC refers to both the EMC-LAM and EMC-LAMx, and NSSL refers to the NSSL-LAM.

deep convection was produced explicitly by the model. The forecast labeled GSL-LAM was initialized with analyses from version 4 of the HRRR and used the forecast from the Rapid Refresh model as LBC. GSL-LAM used similar physics and vertical resolution as EMC-LAMx, except that the Rapid Update Cycle (Smirnova et al., 2016) land surface model was used instead of Noah and there were 64 vertical levels matching the old GFS spectral model configuration instead of 60 matching the NAM model configuration (Figure 1). The differences between GSL-LAM and EMC-LAMx can thus originate from the initial and LBC, the differences in vertical levels, and/or the different land surface model. Rather than controlling for a single major difference in FV3 configuration, the inclusion of GSL-LAM allows for quantification of the impacts of several configuration changes that are less directly linked to the expected performance of the simulated radiance forecasts.

Forecast simulated radiances were calculated hourly using the Community Radiative Transfer Model (Y. Han et al., 2006) version 2.3.0 (CRTM, 2019). For the EMC and NSSL forecasts, this was done within the Unified Post-Processor (Wolff et al., 2016) version 4.1.0 (UPP, 2019, https://dtcenter.org/sites/default/files/community-code/upp-users-guide-v4.1.pdf, accessed 25 January 2023). For the GSL forecasts, this was done within the Gridpoint Statistical Interpolation (Shao et al., 2016) version 3.7 (GSI, 2020).

2.2. Verification Metrics

The corresponding hourly ABI radiance observations are obtained from the National Centers for Environmental Information (GOES-R, 2020) data repository at https://www.ncdc.noaa.gov/airs-web/search. This evaluation focuses on the water-vapor sensitive channels 8 (6.2 micron), 9 (6.9 micron) and 10 (7.3 micron), which have clear-air maximum weighting functions at about 325, 400, and 625 hPa, respectively. Grid-point based diagnostics such as root-mean-square error (RMSE) and bias were calculated after interpolating the ABI radiance observations from their native ~2 km grid to the model grid using bi-linear interpolation. Neighborhood-based and object-based diagnostics are also calculated after first interpolating the observed radiances to the model grid.

Grid-point based verification can be influenced by a "double-penalty" effect (e.g., Baldwin et al., 2001; Johnson, Wang, Kong, & Xue, 2011) in the presence of high amplitude features such as clouds because small spatial errors can result in improvements to the objective verification metric by simply forecasting less clouds overall (i.e., low bias). The Fractions Skill Score (FSS; Gasperoni et al., 2020; Roberts & Lean, 2008; Schwartz et al., 2010) can mitigate the double-penalty effect while also providing verification metrics at different brightness temperature thresholds. The FSS is a skill score calculated as $FSS = 1 - FS/FS_{ref}$, where FS is the Fractions Score and FS_{ref} is a reference Fractions Score. The Fractions Score is the root-mean-square difference between the fraction of neighboring gridpoints forecast to exceed some threshold and the fraction of neighboring grid points observed to exceed the same threshold, calculated for each grid point in turn. Here, a circular neighborhood is used defined by a neighborhood radius. The FS_{ref} is the FS that would result if there was zero overlap between the forecast and observed features and is calculated as the sum of mean-squared forecast and observed neighborhood fractions. An advantage of the FSS is that it can evaluate the forecast skill with respect to different intensity (threshold) and spatial scale (radius) of features. FSS is herein calculated at intensity thresholds of 210, 220, and 230 K and an 18 km radius. The 18 km radius is chosen to remove the effects of spatial displacements that are below the model effective resolution of approximately 7 times the model grid spacing (Skamarock, 2004). FSS with larger radii showed similar relative performance of the forecasts as the 18 km radius (not shown), suggesting that mesoscale displacement of otherwise well-forecast features was not the primary cause of differences in forecast skill. The intensity thresholds are chosen to be sensitive to the coldest cloud tops in convective cores, high clouds generally associated with deep convection, and more general clouds that may have tops at the lower altitude 230 K level. An example of channel 10 forecast differences for the 14 May 2020 case at 24-hr lead time is shown in Figure 2.

Object-based diagnostics are obtained using the Development Testbed Center Method for Object-based Diagnostic Evaluation component of the METplus (Brown et al., 2021) software (Davis et al., 2006, 2009). The object-based evaluation consists of identification of objects in forecast and observed fields, as well as matching
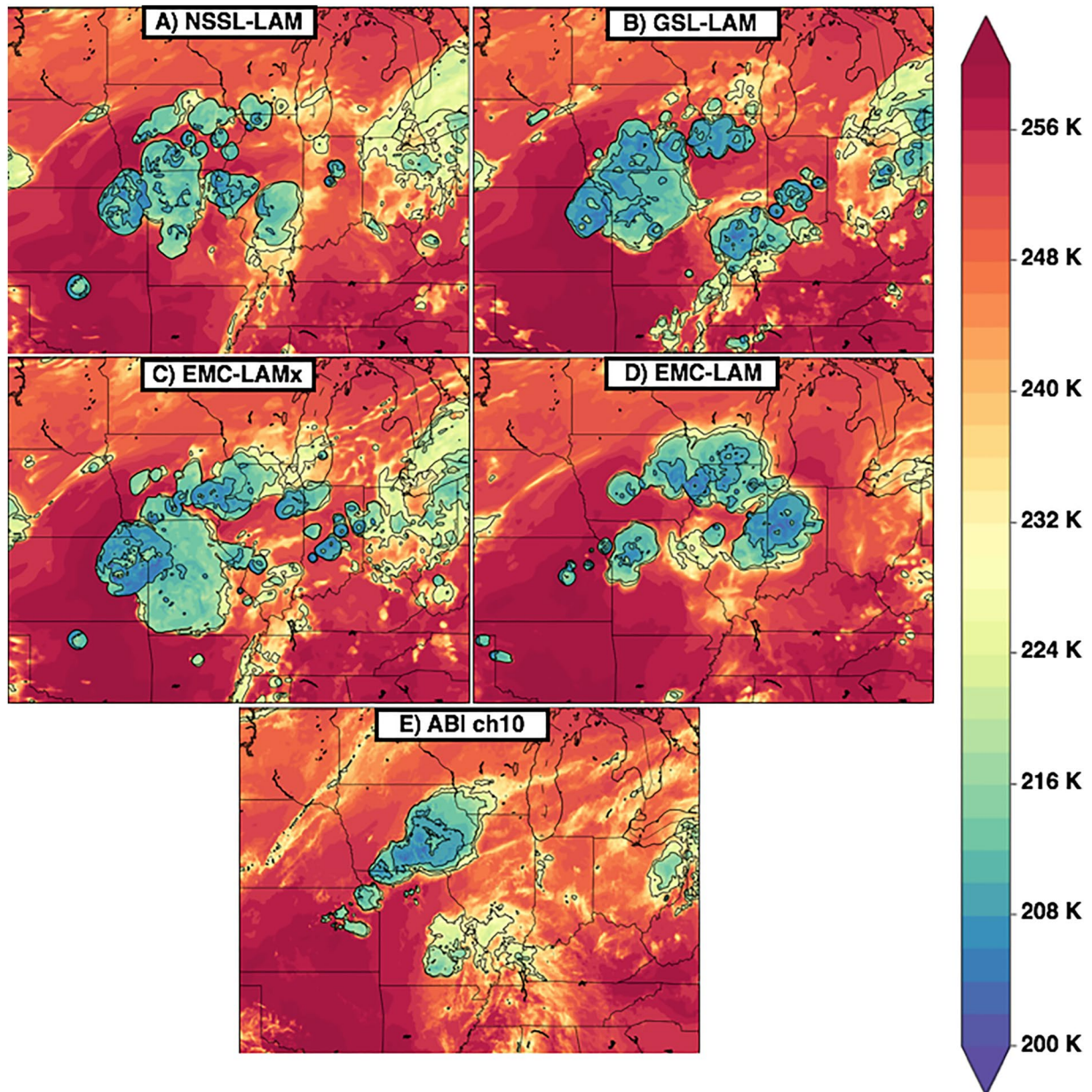
Figure 2. Example of ABI Channel 10 (a) NSSL-LAM simulated radiance, (b) GSL-LAM simulated radiance, (c) EMC-LAMx simulated radiance, (d) EMC-LAM simulated radiance, and (e) observed radiance, valid at 0000 UTC 15 May with forecasts at the 24-hr lead time. Contours (black) are overlaid at the 210, 220, and 230 K brightness temperature thresholds.

of forecast and observed objects. An advantage of the object-based framework is that it allows for an objective verification against observations in terms of object attributes such as size, shape and location rather than simple spatial overlap that grid-point and neighborhood-based frameworks are suited to evaluate. In the present study, cloud objects are identified using a threshold of 230 K applied to the channel 10 observed and simulated radiances, after applying a smoother that consists of averaging each grid point within a 20 km radius. The 230 K threshold is selected to include both deep convective and lower-altitude cloud systems (e.g., Figure 2). While the 220 K threshold objects (not shown) yields many similar conclusions as with the 230 K threshold, reducing the threshold results in noisier statistics due to the already-reduced sample size when moving from the grid-point space to the object space, and are therefore not included.
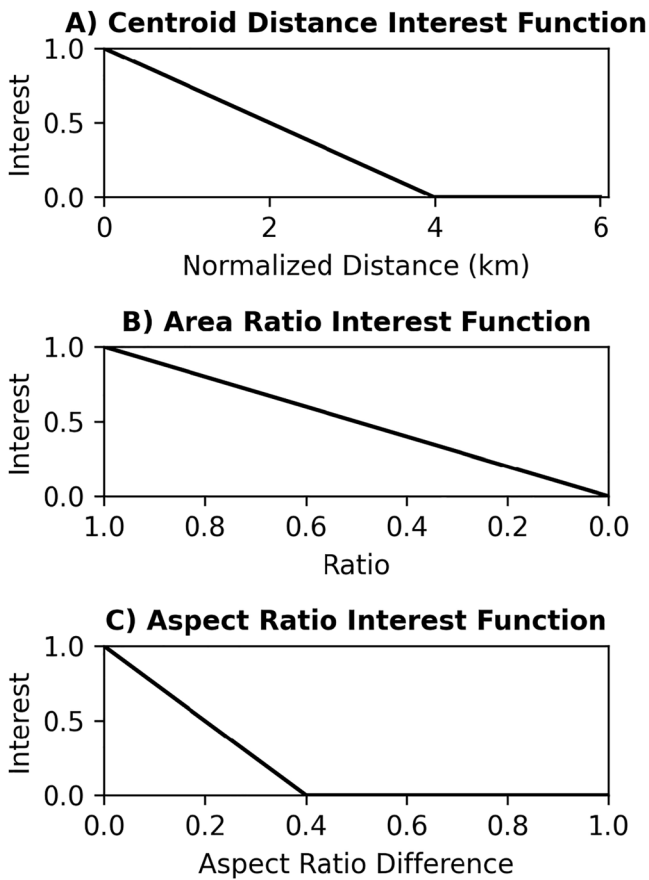
A) Centroid Distance Interest Function

B) Area Ratio Interest Function

C) Aspect Ratio Interest Function

Figure 3. Interest functions used for calculating similarity of matched objects, for attributes of (a) centroid distance, (b) area ratio, and (c) aspect ratio difference.

The Object-based Threat Score (OTS) is calculated as the ratio of the area of forecast and observed matched object pairs, weighted by the similarity of the matched objects, normalized by the total area of all forecast and observed objects (Johnson, Wang, Xue, & Kong, 2011). The similarity of matched objects is calculated using Total Interest, which is a weighted average of the interest, or similarity, of different attributes of the paired objects. Here, the attributes of centroid distance, area ratio and aspect ratio difference are used. In the weighted average, a weight of 1.0 is used for centroid distance while the weight for area ratio and aspect ratio difference is equal to the centroid distance interest value. Thus, objects that are spatially distant have less contribution from size and shape similarities than nearby objects which are more likely to represent the same cloud system. The interest of the individual attributes is calculated as shown in Figure 3. Importantly, several different functions for the individual attribute interests and several different combinations of weights were also tried and the relative performance of the different forecasts was in general not sensitive to the details of the matching function.

Statistical significance of differences between the forecast performance metrics (RMSE, bias, FSS and OTS) is tested using a permutation resampling method following Hamill (1999). In short, a distribution of 1,000 differences between a metric (e.g., RMSE) is obtained by randomly considering one of the forecasts being compared as "forecast 1" and the other as "forecast 2" separately on each forecast case. The distribution of differences in the metric for the randomly assigned forecasts is then used to determine the probability that the true difference between the two forecasts is smaller than the difference that can occur by random chance (i.e., a p-value). The p-value is thus the percentage of randomly assigned difference magnitudes that exceed the magnitude of the actual difference between the two forecasts being compared. Statistical significance at the 95% confidence level (i.e., p-value < 0.05) is indicated on the corresponding figures with a black dot for the difference between NSSL-LAM and EMC-LAMX, a blue dot for the difference between EMC-LAM and EMC-LAMx, and an orange dot for the difference between GSL-LAM and EMC-LAMx.

3. Results

3.1. Impacts of Physics Configuration and Vertical Levels

The overall bias and RMSE of the simulated ABI radiance forecasts are shown in Figure 4. The biases (i.e., average difference between simulated and observed radiance) are generally on the order of 1–2 K, while the RMSE ranges from ~5 K for channel 8–~10 K for channel 10. The relative biases among the forecasts are similar across all three channels, with EMC-LAM being the warmest and EMC-LAMx being the coldest. The least biased forecast (i.e., closest to zero bias) depends on the channel and forecast lead time within the diurnal cycle. At early lead times, EMC-LAMx is generally least biased and EMC-LAM is generally most biased. At the 1-day lead time, during the diurnal maximum of convection, EMC-LAM is least biased for channel 9 while NSSL-LAM is generally least biased for channels 8 and 10. EMC-LAM consistently has the smallest RMSE while EMC-LAMx and GSL-LAM consistently have the largest RMSE. The larger differences in bias and RMSE between EMC-LAM and the other three forecasts than the differences in bias and RMSE among the other three forecasts suggest that the physics configuration has a larger impact on the grid-point based verification of the ABI water vapor sensitive radiances than the vertical resolution and source of IC. However, it is also notable that there are consistent differences between NSSL-LAM and EMC-LAMx which suggests that the vertical resolution does also have an impact on both the bias and RMSE of the simulated radiance forecasts. It is also noted that the difference in RMSE between EMC-LAMx and GSL-LAM are much smaller than the differences between EMC-LAMx and NSSL-LAM or EMC-LAM. This suggests that the impact on simulated radiance forecast performance of changing vertical resolution or the microphysics and PBL scheme were greater than the impact of changing the source
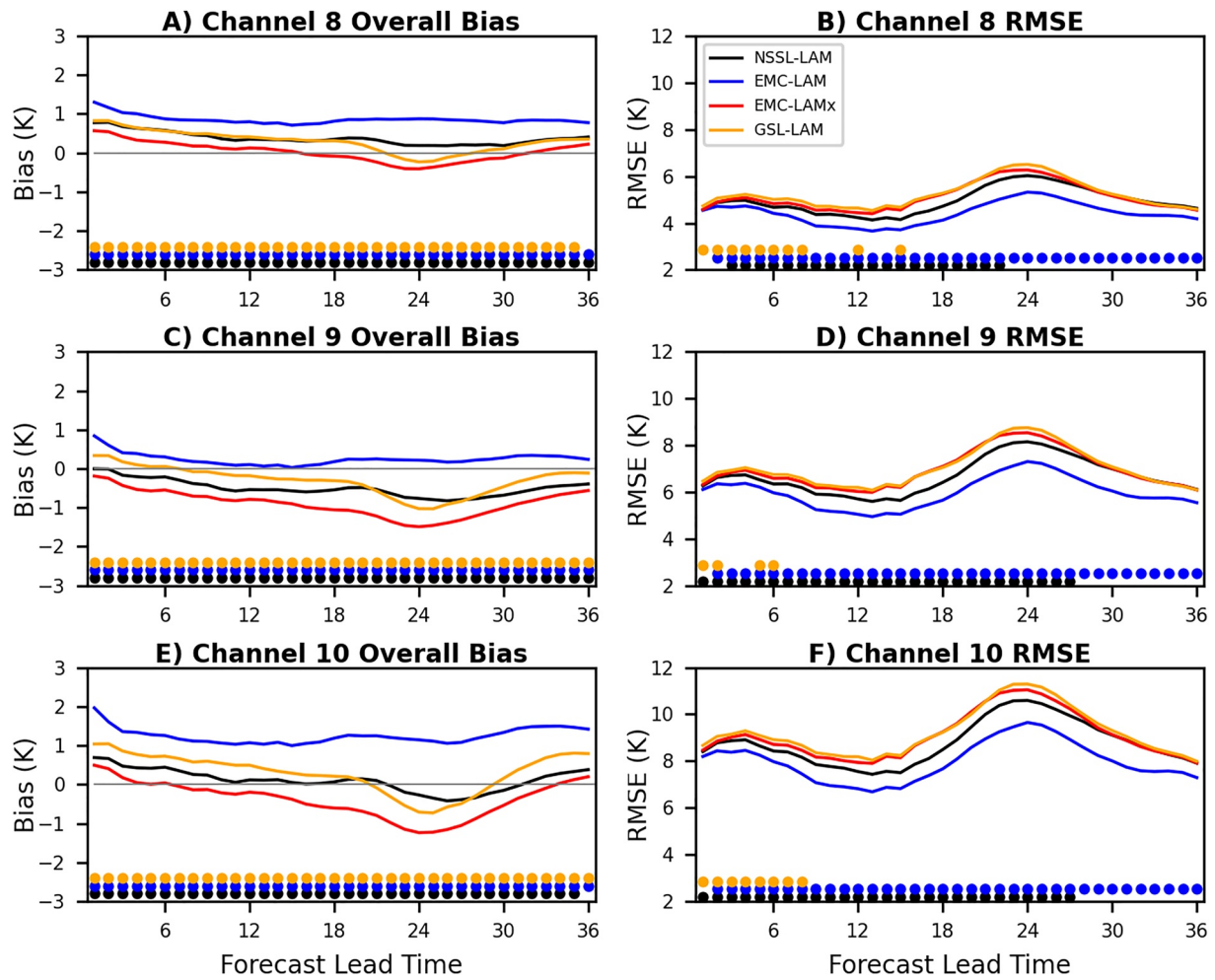
Figure 4. Grid-point verification over all 14 cases for (a, c, e) bias and (b, d, f) RMSE of (a), (b) channel 8, (c), (d) channel 9 and (e), (f) channel 10 simulated radiance forecasts. Markers along the horizontal axis indicate statistically significant difference at the 95% confidence level for (black) NSSL-LAM versus EMC-LAMx, (blue) EMC-LAM versus EMC-LAMx, and (orange) GSL-LAM versus EMC-LAMx.

of IC, land surface model, and the distribution of a similar number of vertical levels. The differences in bias are statistically significant at the 95% level for almost all lead times. The RMSE differences between EMC-LAM and EMC-LAMx are all statistically significant except at the 1-hr lead time. The differences between NSSL-LAM and EMC-LAMx are significant at most lead times. The differences between GSL-LAM and EMC-LAMx are most consistently significant during the first 6 forecast hours which suggests the most consistent differences are largely due to the different initial conditions (Figure 4).

The overall warmer brightness temperatures for EMC-LAM than the other forecasts seen in Figures 4a, 4c, and 4e suggest that the EMC-LAM may be forecasting fewer overall clouds with cold brightness temperature than the other forecasts. This interpretation is confirmed by using the channel 10 minus channel 8 brightness temperature difference as an approximate classification of pixels with deep clouds (small difference between channel 8 and 10), mid-level clouds (moderate difference between channel 8 and 10) and clear air or very low-level clouds (large difference between channel 8 and 10), following the classification used by Cintineo et al. (2014). All of the FV3-LAM configurations over-forecast the coverage of deep clouds compared to observations at most lead times (Figure 5a). This over-forecasting is generally least pronounced for EMC-LAM, which is likely related to both the presence of cumulus parameterization in EMC-LAM which can mitigate the over-forecasting of deep convection, as well as the more expansive anvils clouds often associated with organized convection using Thompson microphysics than other microphysics schemes (e.g., Feng et al., 2018; Wheatley et al., 2014). NSSL-LAM also has less over-forecasting of high clouds than EMC-LAMx at lead times of ~4–24 hr, showing that the increased
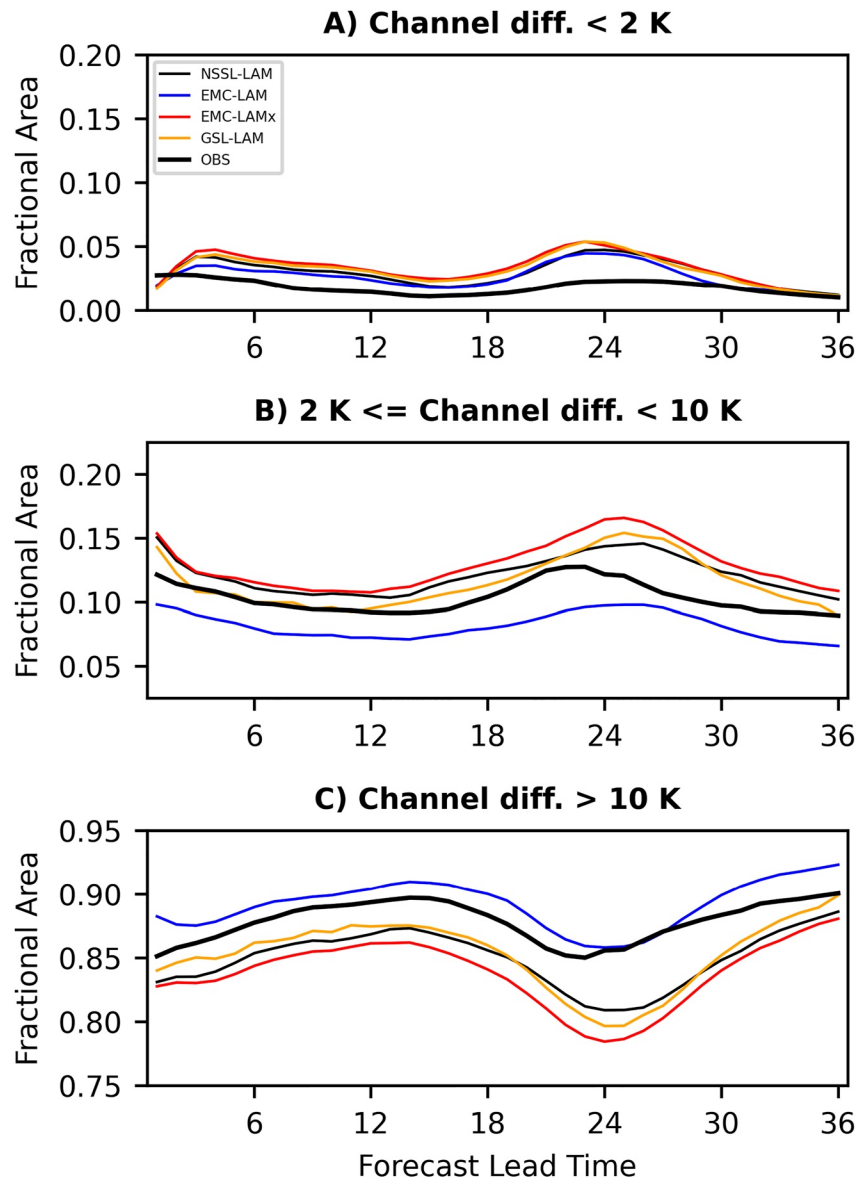
Figure 5. Fraction of domain with channel 8–channel 10 brightness temperature difference (a) less than 2 K, (b) between 2 and 10 K, and (c) greater than 10 K, calculated over all 14 cases.

number of vertical levels also helps mitigate the over-forecasting of high clouds that are often associated with deep convection.

For the mid-level clouds (Figure 5b), all FV3-LAM configurations show a late timing bias in the diurnal maximum compared to observations. While EMC-LAM under-forecasted the coverage of mid-level clouds, the other three configurations over-forecasted the coverage of mid-level clouds. The impact of vertical resolution (NSSL-LAM vs. EMC-LAMx) on reducing the over-forecasting of mid-level clouds is most pronounced during and after the diurnal convective maximum. The EMC-LAM slightly over-forecasts the coverage of clear-sky pixels (Figure 5c), while the other three configurations under-forecast the coverage of clear-sky pixels, consistent with the trends seen in Figures 5a and 5b. The coverage of deep clouds for EMC-LAMx and GSL-LAM are very similar (Figure 5a), consistent with the similar RMSE for these forecasts in Figure 4. However, the differences in clear air (Figure 5c) and mid-level (Figure 5b) clouds between EMC-LAMx and GSL-LAM is generally similar in magnitude to the differences between EMC-LAMx and NSSL-LAM.
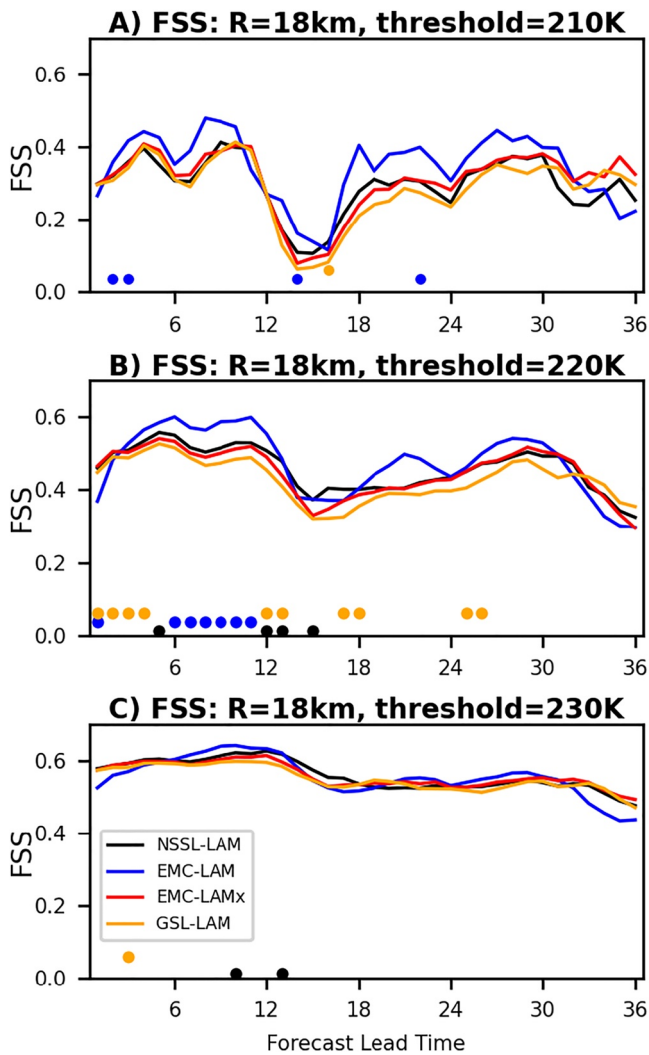
Figure 6. Fractions Skill Score of ABI channel 10 simulated radiance over all 14 cases as a function of forecast lead time with a radius of 18 km and radiance thresholds below (a) 210 K, (b) 220 K, and (c) 230 K. Statistical significance is indicated as in Figure 4.

The FSS for channel 10 radiance verification at three brightness temperature thresholds and an 18 km neighborhood radius is shown in Figure 6. Other neighborhood radii show very similar relative performance among the forecasts (not shown). The FSS for the coldest cloud tops below 210 K (Figure 6a) is generally higher for EMC-LAM than the other three configurations, except for the first forecast hour and some of the lead times during the minimum in the diurnal cycle of deep convection around forecast hours 12–17 and 32–36. This advantage of EMC-LAM at many lead times is also seen at the 220 K threshold (Figure 6b). However, for the warmer threshold of 230 K (Figure 6c), which also includes mid-level clouds, the performance among the different configurations is more similar than at the colder thresholds, except for the EMC-LAM under-performance at very early, and very late, forecast lead times. There is a small advantage of the increased vertical levels in NSSL-LAM, compared to EMC-LAMx, at several lead times and thresholds (Figure 6). However, the difference is less pronounced than in the grid-point based verification and is comparable in magnitude to the differences between EMC-LAMx and GSL-LAM. While many of these differences are statistically significant at the 95% level, especially at the 220 K threshold, there are many times for which high levels of statistical significance of FSS differences are not obtained in the 14-case experiment.

While the above grid-point and neighborhood-based diagnostics can quantify the relative performance of different FV3-LAM configurations, object-based verification can provide an additional perspective by relating the performance differences to specific cloud object attributes. Figure 7 shows average size, total number, and size distribution of cloud objects using a 230 K threshold on the channel 10 radiance to identify cloud objects. All configurations have an average cloud object size that is too small, especially EMC-LAM (Figure 7b). However, the total number of cloud objects is most consistent with observations for EMC-LAM, while the other configurations over-forecast the number of cloud objects at most lead times (Figure 7a). The under-forecasting of clouds in EMC-LAM inferred from Figure 5 occurs for both small cloud objects (Figure 7c) and large cloud objects (Figure 7d). The other three configurations generally over-forecast the number of both small (<2,000 km^2) and very large (>20,000 km^2) cloud objects. These trends in over- and under-forecasting of clouds are most consistent with the mid-level clouds (Figure 5b) rather than the high-level clouds (Figure 5a).

The object-based framework also provides an additional perspective on the skill of the forecasts of simulated cloud features. EMC-LAM has the best forecasts of simulated clouds at about the 10–32 hr forecast lead times in the object-based framework (Figure 8). However, the EMC-LAM forecast is relatively poor compared to NSSL-LAM, and similar to EMC-LAMx at early lead times, consistent with the reduced performance of EMC-LAM at ~1–3 hr lead times seen in Figure 6. The greater number of vertical levels in NSSL-LAM compared to EMC-LAMx generally results in better simulated cloud forecasts, especially at early lead times and during the diurnal convection maximum (Figure 8). Most of the statistically significant differences in OTS occur during the diurnal convective maximum of ~23–31 hr lead times. The lack of high statistical significance at many lead times is due to only having 14 cases of controlled experiments, and the enhanced variability resulting from having fewer objects in each forecast than the number of grid points used for calculated other metrics such as RMSE. However, consistency of differences such as EMC-LAM versus EMC-LAMx across multiple lead times and verification metrics provides confidence in the results that is not necessarily reflected in the statistical significance tests.
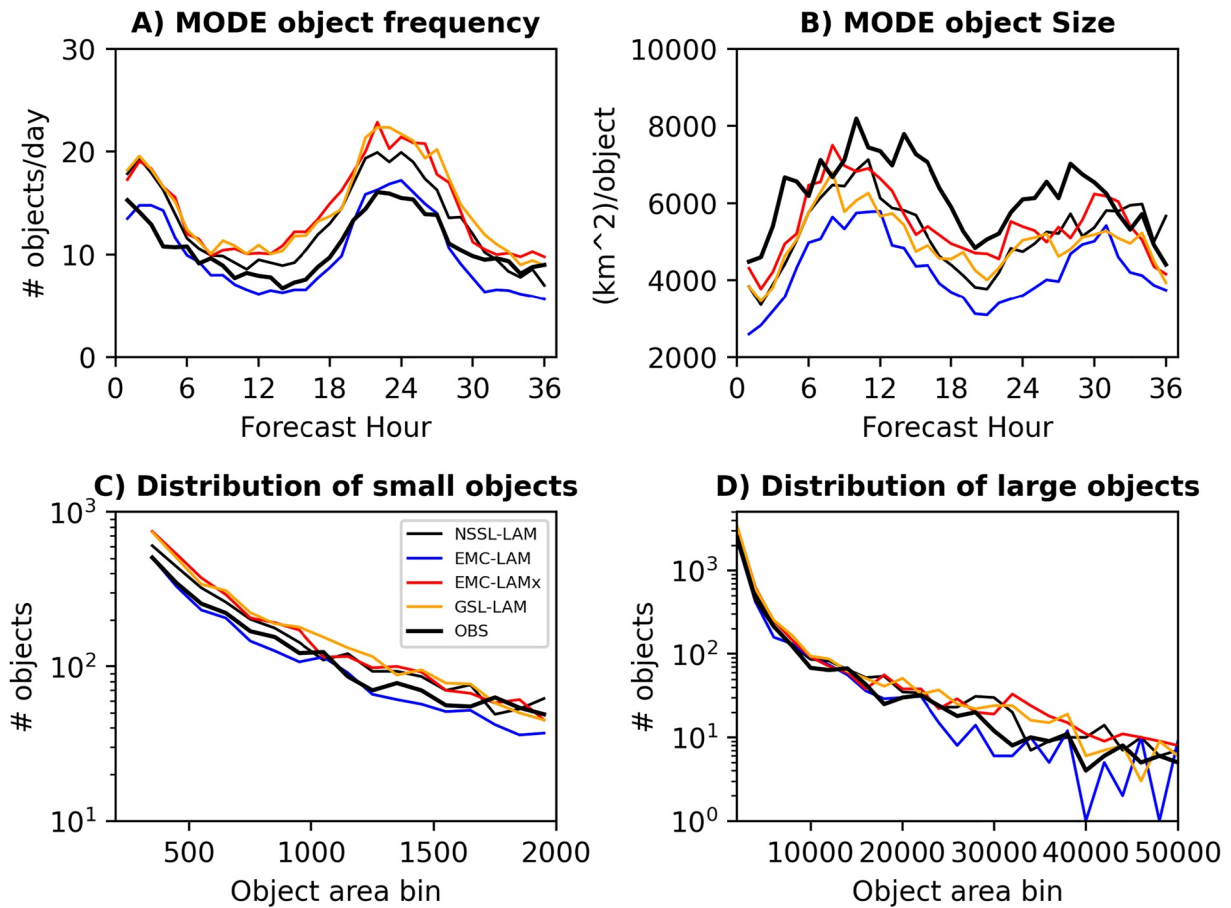
Figure 7. Object size attributes, including (a) average number of cloud objects per day as a function of forecast lead time, (b) average cloud object size as a function of forecast lead time, (c) average number of objects in size bins up to 2,000 km^2 at all forecast lead times, and (d) average number of objects in size bins above 2,000 km^2 at all forecast lead times.

3.2. Impacts of Bias Correction

Similar to Chandramouli et al. (2022), the bias is calculated using a binning approach as a function of symmetric brightness temperature, which is the average of the forecast and observed brightness temperature. The estimated bias is therefore non-linear in that the bias is neither constant nor a linear function of the brightness temperature. The bias is plotted as a function of symmetric brightness temperature rather than only forecast or only observed brightness temperature because bias correction as a function of only observed or only forecast brightness temperature yields subjectively unrealistic bias-corrected brightness temperature (not shown) as a result of the biases in cloud cover discussed in Section 3.1. The bias of brightness temperature is generally consistent across forecast lead times and is shown in Figure 9 for the representative example of the 24-hr lead time. The forecasts with Thompson microphysics and MYNN PBL scheme (NSSL-LAM, EMC-LAMx, and GSL-LAM) have a more pronounced negative bias, largely due to over-forecasting high-level clouds, than EMC-LAM at symmetric brightness temperatures between about 210 and 230 K (Figure 9). Compared to EMC-LAMx with fewer vertical levels, NSSL-LAM has less bias at ~230 K but more bias at ~220 K. For mid-level clouds with brightness temperature ~230–240 K, EMC-LAM has a positive bias consistent with too few forecast clouds, while the other configurations have a negative bias. The bias distributions for EMC-LAMx and GSL-LAM are again more similar to
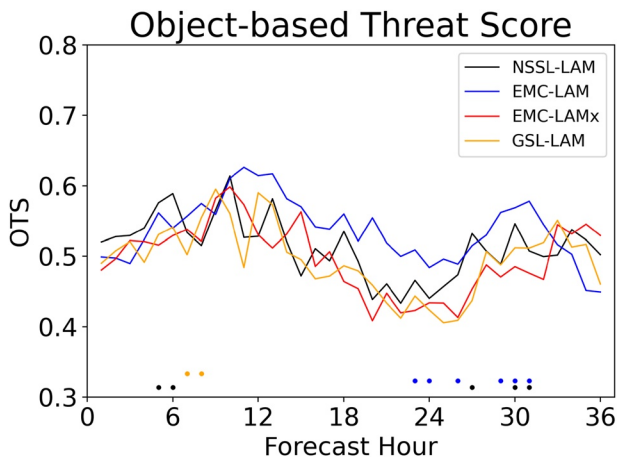

Figure 8. Object-based Threat Score as a function of forecast lead time. Statistical significance is indicated as in Figure 4.
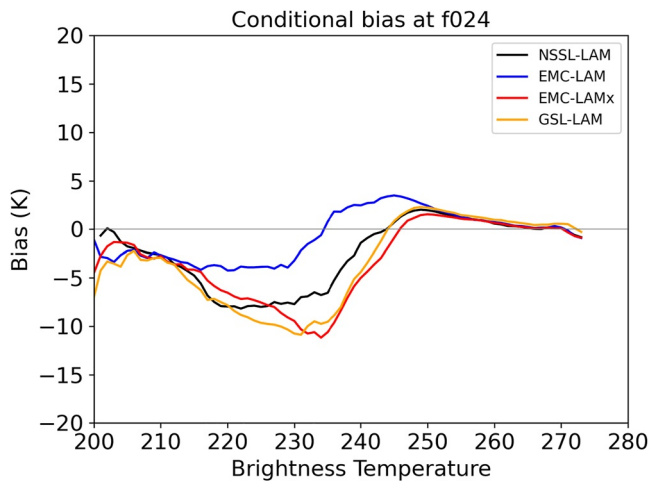
Figure 9. Systematic bias as a function of symmetric brightness temperature.

each other than to the other forecasts (Figure 9). Overall, the FV3 configuration is confirmed to affect the non-linear bias characteristics of simulated radiance besides just the overall bias and skill discussed above.

The RMSE of the simulated radiance forecasts after removing the non-linear biases shown in Figure 9 is shown in Figure 10. After removing the predictable part of the error (i.e., bias), the RMSE in general is reduced from about 5 to 10 K (Figure 4) to about 2–2.5 K (Figure 10). After bias correction, EMC-LAM has the lowest RMSE among the considered forecasts for all three channels, although the difference is reduced compared to Figure 4. The increased vertical levels of NSSL-LAM, compared to EMC-LAMx, slightly reduces the RMSE during the mid-day and afternoon hours of ~12–24 hr forecast lead times. The differences in RMSE among NSSL-LAM, EMC-LAMx, and GSL-LAM after bias correction are ~0.1 K or smaller (Figure 10). However, the statistical significance of the differences is similar to the differences in RMSE without bias correction from Figure 4.

From a data assimilation perspective, the agreement between the vertical structure of observed and simulated clouds is an important consideration. This is because assimilated brightness temperature observations would be expected to better maintain their influence into the forecast period when there is consistency between the cloud structures that the model can support and those that are observed. Therefore, the channel 10-8 brightness temperature difference, which is sensitive to cloud vertical structure (Cintineo et al., 2014), is also considered before and after bias correction (Figure 11). Before bias correction, NSSL-LAM, EMC-LAMx, and GSL-LAM have similar distributions of the channel 10-8 difference, while EMC-LAM has generally too small channel 10-8 difference except in clear air (Figure 11). The smaller channel differences for EMC-LAM likely indicate clouds that tend to be too deep compared to observations, while the negative channel differences may indicate too many overshooting tops in thunderstorms that extend into the warmer stratosphere (Cintineo et al., 2014). The EMC-LAM errors in cloud vertical structure are generally improved with bias correction (Figure 11b). However, all configurations have slightly too large channel differences in deep clouds after bias correction (Figure 11b).

4. Summary and Conclusions

The FV3-based RRFS is planned to replace the WRF-based HRRR as the operational convection-allowing forecast system at NCEP. Evaluations so far of different FV3 configurations for this purpose have largely focused on the precipitating cores of intense convection in terms of reflectivity, updraft helicity and surrogate severe hazards. Non-precipitating cloud regions are also of interest because of their radiative impacts on convective environments and their relationship to the latent heating and microphysical structure of cloud systems. Furthermore, effective assimilation of the high-resolution all-sky infrared radiances provided by the ABI instrument onboard the geostationary GOES-16 and GOES-18 satellites requires an understanding of the forecast error and bias characteristics of simulated radiance and cloud features. The present study addresses this need by systematically evaluating a series of convection-allowing forecasts with different FV3 configurations in terms of their simulated radiances. The controlled experiments (EMC-LAMx, NSSL-LAM, and EMC-LAM) run on the same forecast cases during the 2020 NOAA HWT allow for the impacts of two different physics configurations and vertical resolution configurations to be isolated. An additional experiment (GSL-LAM) further allows for the quantification of the impact of simultaneous changes to the source of initial conditions, land surface model, and distribution of a similar number of vertical levels. The cloudy radiance verification in particular is emphasized through a variety of grid-point, neighborhood-based and object-based verification metrics.

We find that the NSSL-LAM, EMC-LAMx, and GSL-LAM configurations all underperform the EMC-LAM configuration by several metrics after the first few forecast hours. This difference in performance can be attributed to the physics configuration since the only difference between the EMC-LAM and EMC-LAMx is the microphysics, cumulus and PBL parameterization schemes. EMC-LAMx, as well as NSSL-LAM and GSL-LAM, uses the Thompson microphysics scheme which is likely largely responsible for the observed over-forecasting of deep and mid-level clouds in these FV3 forecasts. Although largely inactive, the presence of a cumulus parameterization scheme in EMC-LAM may also contribute to reducing an over-convective bias of the FV3-model. In other
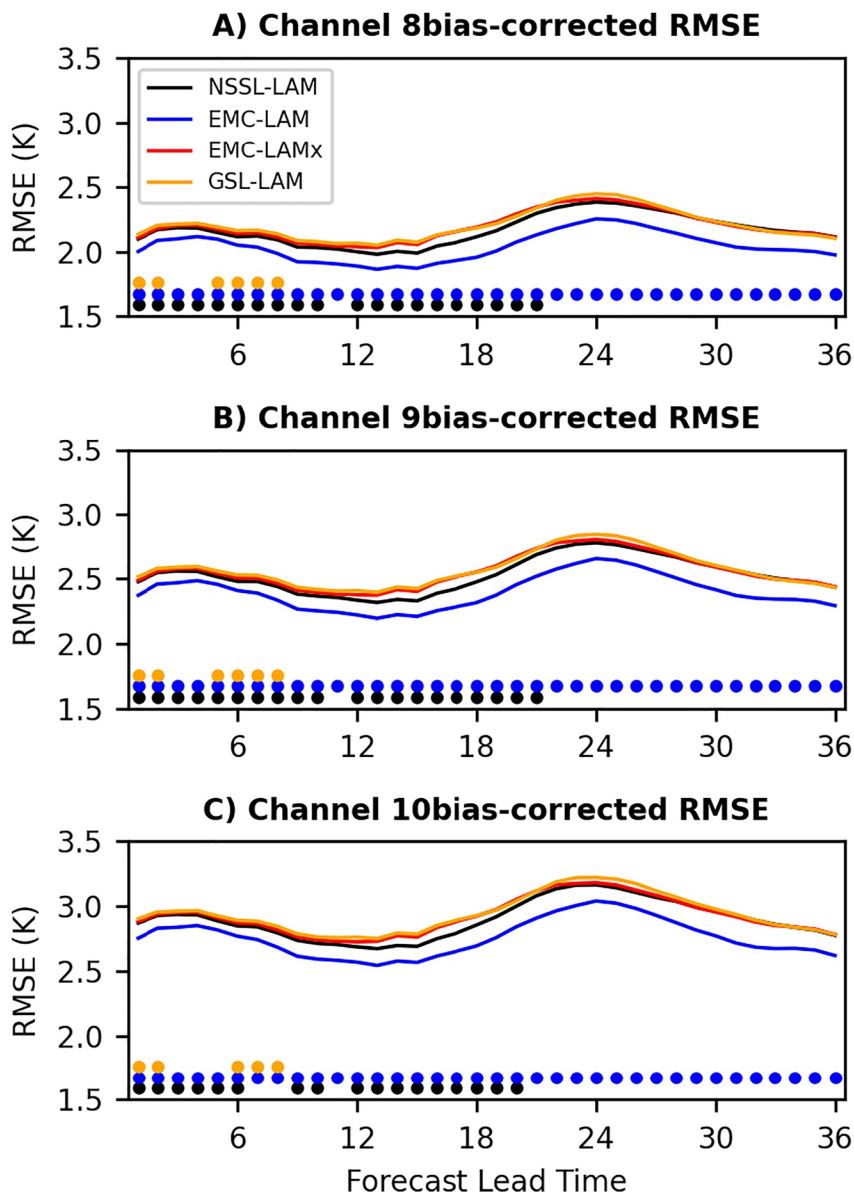
A) Channel 8bias-corrected RMSE

B) Channel 9bias-corrected RMSE

C) Channel 10bias-corrected RMSE

Figure 10. RMSE of simulated (a) channel 8, (b) channel 9, and (c) channel 10 radiances after removing the bias as a function of symmetric brightness temperature. Statistical significance is indicated as in Figure 4.
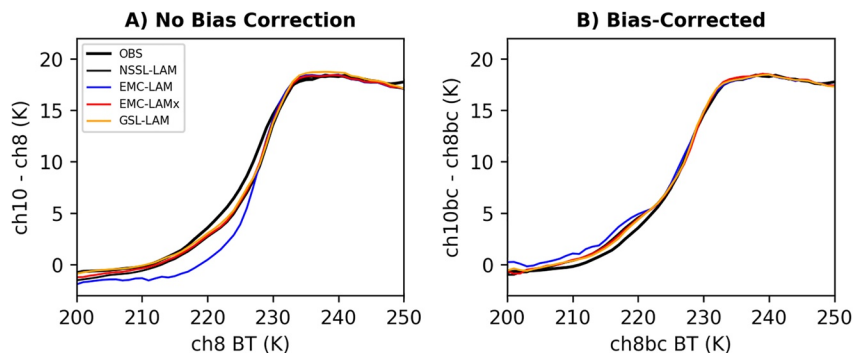
A) No Bias Correction

B) Bias-Corrected

Figure 11. Difference between channel 10 and channel 8 brightness temperature as a function of the channel 8 brightness temperature (a) without and (b) with bias correction.

words, there may be an advantage in retaining some limited cumulus parameterization at 3-km grid spacing in FV3-based forecasts. The over-forecasting of clouds in EMC-LAMx is also slightly improved in the NSSL-LAM configuration, which can be attributed to increased vertical resolution in NSSL-LAM. Since the greatest difference in model layer thicknesses between NSSL-LAM and EMC-LAM occur in mid-levels (Figure 1), we speculate that this impact of vertical resolution may be related to the impacts of mid-level vertical resolution on entrainment processes during deep convection. For most results in this study, the differences between EMC-LAMx and GSL-LAM were much smaller than the differences between EMC-LAMx and NSSL-LAM or EMC-LAM. This relative similarity further supports the conclusion that the microphysics and PBL physics, and to a lesser extent the vertical resolution, are two of the critical aspects of the FV3 configuration from the perspective of forecasting simulated infrared radiances.

It should be noted that the FV3-LAM has undergone significant additional development in dynamics, physics (including Thompson microphysics), and data assimilation since the 2020 HWT. While these developments have shown substantial improvements in terms of radar reflectivity and precipitation since 2020 HWT (not shown), the present study highlights the importance of evaluating the impact of such developments from multiple perspectives, including the perspective of simulated satellite brightness temperatures.

For the purpose of data assimilation, the early lead times and the cloud vertical structure are of particular importance. While the NSSL-LAM, EMC-LAMx, and GSL-LAM, with MYNN PBL and Thompson microphysics, have too much cloud through most of the forecast, the EMC-LAM, with EDMF PBL, and GFDL microphysics with SAS cumulus parameterization, has particularly poor performance at early lead times. Before applying bias-correction, EMC-LAM also shows unrealistic vertical structure of clouds that could be potentially important for the assimilation of multiple water-vapor channels. Bias-correction reduces the error in the EMC-LAM cloud vertical structure, and reduces but does not eliminate the EMC-LAM advantage in terms of RMSE. In general, non-linear bias correction conditioned on symmetric brightness temperature reduces the RMSE of the simulated radiance forecasts by about a factor of 2 for all configurations considered.

Data Availability Statement

The FV3-LAM software (FV3-LAM, 2020) version 1.01 used to run the forecasts evaluated in this study is available at https://github.com/ufs-community/ufs-srweather-app/releases, thanks to the UFS community. The level 1b ABI radiance observations (GOES-R, 2020) are available at https://www.ncdc.noaa.gov/airs-web/search, thanks to NCEP. The CRTM version 2.3.0 (CRTM, 2019) is available at github.com/BenjaminTJohnson/CRTM_REL-2.3.0, the GSI version 3.7 (GSI, 2020) is available at https://dtcenter.org/community-code/gridpoint-statistical-interpolation-gsi/download, and the UPP version 4.1.0 (UPP, 2019) is available at https://github.com/NOAA-EMC/UPP/commit/3b42c2a4999283bce18f7a45bccf4610703ad556. The simulated radiance forecasts and associated metadata can be accessed at http://weather.ou.edu/~map/datasets/johnsonetal2023.tar.

References

Aligo, E. A., Ferrier, B., & Carley, J. R. (2018). Modified NAM microphysics for forecasts of deep convective storms. *Monthly Weather Review*, *146*(12), 4115–4153. https://doi.org/10.1175/Mwr-D-17-0277.1

Aligo, E. A., Gallus, W. A., Jr., & Segal, M. (2009). On the impact of WRF model vertical grid resolution on midwest summer rainfall forecasts. *Weather and Forecasting*, *24*(2), 575–594. https://doi.org/10.1175/2008WAF2007101.1

Baldwin, M. E., Lakshmivarahan, S., & Kain, J. S. (2001). Verification of mesoscale features in NWP models. In *Ninth conference on mesoscale processes* (pp. 255–258). American Meteorological Society.

Barlage, M., Chen, F., Tewari, M., Ikeda, K., Gochis, D., Dudhia, J., et al. (2010). Noah land surface model modifications to improve snowpack prediction in the Colorado Rocky Mountains. *Journal of Geophysical Research*, *115*(D22), D22101. https://doi.org/10.1029/2009JD013470

Black, T. L., Abeles, J. A., Blake, B. T., Jovic, D., Rogers, E., Zhang, X., et al. (2021). A limited area modeling capability for the finite-volume cubed-sphere (FV3) dynamical core and comparison with a global two-way nest. *Journal of Advances in Modeling Earth Systems*, *13*(6), e2021MS002483. https://doi.org/10.1029/2021MS002483

Brown, B., Jensen, T., Gotway, J. H., Bullock, R., Gilleland, E., Fowler, T., et al. (2021). The Model Evaluation Tools (MET): More than a decade of community-supported forecast verification. *BAMS*, *102*(4), E782–E807. https://doi.org/10.1175/BAMS-D-19-0093.1

Carley, J. R., Alexander, C. R., Wolff, J. K., Beck, J., Wicker, L. J., Rogers, E., et al. (2021). Status of NOAA's next generation convection-allowing ensemble: The rapid refresh forecast system. Special symposium on global and mesoscale models: Updates and center overviews. In *AMS virtual meeting*. Retrieved from https://ams.confex.com/ams/101ANNUAL/meetingapp.cgi/Paper/378383

Chandramouli, K., Wang, X., Johnson, A., & Otkin, J. (2022). Online nonlinear bias correction in ensemble Kalman filter to assimilate GOES-R all-sky radiances for the analysis and prediction of rapidly developing supercells. *Journal of Advances in Modeling Earth Systems*, *14*(3), e2021MS002711. https://doi.org/10.1029/2021MS002711

Chen, J.-H., & Lin, S.-J. (2013). Seasonal predictions of tropical cyclones using a 25-km-resolution general circulation model. *Journal of Climate*, *26*(2), 380–398. https://doi.org/10.1175/JCLI-D-12-00061.1

Cintineo, R., Otkin, J. A., Xue, M., & Kong, F. (2014). Evaluating the performance of planetary boundary layer and cloud microphysical parameterization schemes in convection-permitting ensemble forecasts using synthetic GOES-13 satellite observations. *Monthly Weather Review*, *142*(1), 163–182. https://doi.org/10.1175/MWR-D-13-00143.1

Clark, A. J., Coniglio, M. C., Coffer, B. E., Thompson, G., Xue, M., & Kong, F. (2015). Sensitivity of 24-h forecast dryline position and structure to boundary layer parameterizations in convection-allowing WRF Model simulations. *Weather and Forecasting*, *30*(3), 613–638. https://doi.org/10.1175/WAF-D-14-00078.1

Clough, S. A., Shephard, M. W., Mlawer, E. J., Delamere, J. S., Iacono, M. J., Cady-Pereira, K., et al. (2005). Atmospheric radiative transfer modeling: A summary of the AER codes. *Journal of Quantitative Spectroscopy & Radiative Transfer*, *91*(2), 233–244. https://doi.org/10.1016/j.jqsrt.2004.05.058

CRTM. (2019). Community radiative transfer model, version 2.3.0. [Software]. Retrieved from github.com/BenjaminTJohnson/CRTM_REL-2.3.0

Davis, C., Brown, B., & Bullock, R. (2006). Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. *Monthly Weather Review*, *134*(7), 1772–1784. https://doi.org/10.1175/MWR3145.1

Davis, C., Brown, B., Bullock, R., & Halley-Gotway, J. (2009). The Method for Object-Based Diagnostic Evaluation (MODE) applied to numerical forecasts from the 2005 NSSL/SPC spring program. *Weather and Forecasting*, *24*(5), 1252–1267. https://doi.org/10.1175/2009WAF2222241.1

Dowell, D. C., Alexander, C. R., James, E. P., Weygandt, S. S., Benjamin, S. G., Manikin, G. S., et al. (2022). The High-Resolution Rapid Refresh (HRRR): An hourly updating convection-allowing forecast model. Part I: Motivation and system description. *Weather and Forecasting*, *37*(8), 1371–1395. https://doi.org/10.1175/WAF-D-21-0151.1

Feng, Z., Leung, L. R., Houze, R. A., Jr., Hagos, S., Hardin, J., Yang, Q., et al. (2018). Structure and evolution of mesoscale convective systems: Sensitivity to cloud microphysics in convection-permitting simulations over the United States. *Journal of Advances in Modeling Earth Systems*, *10*(7), 1470–1494. https://doi.org/10.1029/2018MS001305

FV3-LAM. (2020). Finite volume cubed sphere limited area model with UFS workflow, version 1.01 [Software]. https://doi.org/10.5281/zenodo.6505853

Gallo, B. T., Wolff, J. K., Clark, A. J., Jirak, I., Blank, L. R., Roberts, B., et al. (2021). Exploring convection-allowing model evaluation strategies for severe local storms using the finite-volume cubed-sphere (FV3) model core. *Weather and Forecasting*, *36*(1), 3–19. https://doi.org/10.1175/WAF-D-20-0090.1

Gasperoni, N. A., Wang, X., & Wang, Y. (2020). A comparison of methods to sample model errors for convection-allowing ensemble forecasts in the setting of multiscale initial conditions produced by the GSI-based EnVar assimilation system. *Monthly Weather Review*, *148*(3), 1177–1203. https://doi.org/10.1175/MWR-D-19-0124.1

GOES-R Calibration Working Group and GOES-R Series Program. (2020). NOAA GOES-R Se-ries advanced baseline imager (ABI) level 1b radiances [Dataset]. NOAA National Centers for Environmental Information. https://doi.org/10.7289/V5BV7DSR

Grasso, L. D., Lindsey, D. T., Sunny Lim, K.-S., Clark, A. J., Bikos, D., & Dembek, S. R. (2014). Evaluation of and suggested improvements to the WSM6 microphysics in WRF-ARW using synthetic and observed GOES-13 imagery. *Monthly Weather Review*, *142*(10), 3635–3650. https://doi.org/10.1175/MWR-D-14-00005.1

Griffin, S. M., Otkin, J. A., Rozoff, C. M., Sieglaff, J. M., Cronce, L. M., Alexander, C. R., et al. (2017). Seasonal analysis of cloud objects in the High-Resolution Rapid Refresh (HRRR) model using object-based verification. *Journal of Applied Meteorology and Climatology*, *56*(8), 2317–2334. https://doi.org/10.1175/JAMC-D-17-0004.1

GSI. (2020). Communty gridpoint statistical interpolation, version 3.7 [Software]. Retrieved from https://dtcenter.org/community-code/gridpoint-statistical-interpolation-gsi/citation-publications

Hamill, T. M. (1999). Hypothesis tests for evaluating numerical precipitation forecasts. *Weather and Forecasting*, *14*(2), 155–167. https://doi.org/10.1175/1520-0434(1999)014%3C0155:HTFENP%3E2.0.CO;2

Han, J., & Pan, H.-L. (2011). Revision of convection and vertical diffusion schemes in the NCEP Global Forecast System. *Weather and Forecasting*, *26*(4), 520–533. https://doi.org/10.1175/WAF-D-10-05038.1

Han, J., Witek, M. L., Teixeira, J., Sun, R., Pan, H.-L., Fletcher, J. K., & Bretherton, C. S. (2016). Implementation in the NCEP GFS of a Hybrid Eddy-Diffusivity Mass-Flux (EDMF) boundary layer parameterization with dissipative heating and modified stable boundary layer mixing. *Weather and Forecasting*, *31*(1), 341–352. https://doi.org/10.1175/waf-d-15-0053.1

Han, Y., van Delst, P., Liu, Q., Weng, F., Yan, B., Treadon, R., & Derber, J. (2006). *Community radiative transfer model (CRTM): Version 1, NOAA technical report* (p. 122). NOAA.

Harris, L., Chen, X., Zhou, L., & Chen, J.-H. (2020). The nonhydrostatic solver of the GFDL finite-volume cubed-sphere dynamical core. In *NOAA technical memorandum OAR GFDL* (pp. 2020–3003). https://doi.org/10.25923/9wdt-4895

James, E. P., Alexander, C. R., Dowell, D. C., Weygandt, S. S., Benjamin, S. G., Manikin, G. S., et al. (2022). The High-Resolution Rapid Refresh (HRRR): An hourly updating convection-allowing forecast model. Part II: Forecast performance. *Weather and Forecasting*, *37*(8), 1397–1417. https://doi.org/10.1175/WAF-D-21-0130.1

Janjić, Z. I., & Gall, R. (2012). Scientific documentation of the NCEP nonhydrostatic multiscale model on the B grid (NMMB). Part 1: Dynamics. NCAR Technical Note (NCAR/TN-489+STR), *74*. https://doi.org/10.5065/D6WH2MZX

Janjić, Z. I., Gerrity, J. P., & Nickovic, S. (2001). An alternative approach to nonhydrostatic modeling. *Monthly Weather Review*, *129*(5), 1164–1178. https://doi.org/10.1175/1520-0493(2001)129%3C1164:AAATNM%3E2.0.CO;2

Johnson, A., & Wang, X. (2017). Design and implementation of a GSI-based convection-allowing ensemble data assimilation and forecast system for the PECAN field experiment. Part I: Optimal configurations for nocturnal convection prediction using retrospective cases. *Weather and Forecasting*, *32*(1), 289–315. https://doi.org/10.1175/WAF-D-16-0102.1

Johnson, A., & Wang, X. (2019). Multicase assessment of the impacts of horizontal and vertical grid spacing, and turbulence closure model, on subkilometer-scale simulations of atmospheric bores during PECAN. *Monthly Weather Review*, *147*(5), 1533–1555. https://doi.org/10.1175/MWR-D-18-0322.1

Johnson, A., Wang, X., & Jones, T. (2022). Impacts of assimilating GOES-16 ABI channels 9 and 10 clear air and cloudy radiance observations with additive inflation and adaptive observation error in GSI-EnKF for a case of rapidly evolving severe supercells. *Journal of Geophysical Research: Atmospheres*, *127*(11), e2021JD036157. https://doi.org/10.1029/2021JD036157

Johnson, A., Wang, X., Kong, F., & Xue, M. (2011). Hierarchical cluster analysis of a convection-allowing ensemble during the Hazardous Weather Testbed 2009 Spring Experiment. Part I: Development of the object-oriented cluster analysis method for precipitation fields. *Monthly Weather Review*, *139*(12), 3673–3693. https://doi.org/10.1175/MWR-D-11-00015.1

Johnson, A., Wang, X., Xue, M., & Kong, F. (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. *Monthly Weather Review*, *139*(12), 3694–3710. https://doi.org/10.1175/MWR-D-11-00016.1

Jones, T. A., Skinner, P., Yussouf, N., Knopfmeier, K., Reinhart, A., Wang, X., et al. (2020). Assimilation of GOES-16 radiances and retrievals into the warn-on-forecast system. *Monthly Weather Review*, *148*(5), 1829–1859. https://doi.org/10.1175/MWR-D-19-0379.1

Kain, J. S., Weiss, S. J., Bright, D. R., Baldwin, M. E., Levit, J. J., Carbin, G. W., et al. (2008). Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Weather and Forecasting*, *23*(5), 931–952. https://doi.org/10.1175/WAF2007106.1

Nakanishi, M., & Niino, H. (2009). Development of an improved turbulence closure model for the atmospheric boundary layer. *Journal of the Meteorological Society of Japan. Ser. II*, *87*(5), 895–912. https://doi.org/10.2151/jmsj.87.895

Otkin, J. A., & Greenwald, T. J. (2008). Comparison of WRF model-simulated and MODIS-derived cloud data. *Monthly Weather Review*, *136*(6), 1957–1970. https://doi.org/10.1175/2007MWR2293.1

Potvin, C. K., Carley, J. R., Clark, A. J., Wicker, L. J., Skinner, P. S., Reinhart, A. E., et al. (2019). Systematic comparison of convection-allowing models during the 2017 NOAA HWT spring forecasting experiment. *Weather and Forecasting*, *34*(5), 1395–1416. https://doi.org/10.1175/WAF-D-19-0056.1

Putman, W. M., & Lin, S.-J. (2007). Finite-volume transport on various cubed-sphere grids. *Journal of Computational Physics*, *227*(1), 55–78. https://doi.org/10.1016/j.jcp.2007.07.022

Rempel, M., Senf, F., & Deneke, H. (2017). Object-based metrics for forecast verification of convective development with geostationary satellite data. *Monthly Weather Review*, *145*(8), 3161–3178. https://doi.org/10.1175/MWR-D-16-0480.1

Roberts, N., & Lean, H. W. (2008). Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, *136*(1), 78–97. https://doi.org/10.1175/2007MWR2123.1

Schmit, T. J., Griffith, P., Gushor, M. M., Daniels, J. M., Goodman, S. J., & Lebair, W. J. (2017). A closer look at the ABI on GOES-R series. *BAMS*, *98*(4), 681–698. https://doi.org/10.1175/BAMS-D-15-00230.1

Schwartz, C. S., Kain, J. S., Weiss, S. J., Xue, M., Bright, D. R., Kong, F., et al. (2010). Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Weather and Forecasting*, *25*(1), 263–280. https://doi.org/10.1175/2009WAF2222267.1

Shao, H., Derber, J., Huang, X.-Y., Hu, M., Newman, K., Stark, D., et al. (2016). Bridging research to operations transitions: Status and plans of community GSI. *BAMS*, *97*(8), 1427–1440. https://doi.org/10.1175/BAMS-D-13-00245.1

Skamarock, W. C. (2004). Evaluating mesoscale NWP models using kinetic energy spectra. *Monthly Weather Review*, *132*(12), 3019–3032. https://doi.org/10.1175/mwr2830.1

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D., Duda, M. G., et al. (2008). A description of the advanced research WRF version 3. NCAR Technical Note NCAR/TN-475+STR. (pp. 113). https://doi.org/10.5065/D68S4MVH

Smirnova, T. G., Brown, J. M., Benjamin, S. G., & Kenyon, J. S. (2016). Modifications to the rapid update cycle land surface model (RUC LSM) available in the Weather Research and Forecasting (WRF) model. *Monthly Weather Review*, *144*(5), 1851–1865. https://doi.org/10.1175/MWR-D-15-0198.1

Snook, N., Kong, F., Brewster, K., Xue, M., Thomas, K. W., Supinie, T. A., et al. (2019). Evaluation of convection-permitting precipitation forecast products using WRF, NMMB, and FV3 for the 2016–17 NOAA hydrometeorology testbed flash flood and intense rainfall experiments. *Weather and Forecasting*, *34*(3), 781–804. https://doi.org/10.1175/WAF-D-18-0155.1

Thompson, G., Field, P. R., Rasmussen, R. M., & Hall, W. D. (2008). Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization. *Monthly Weather Review*, *136*(12), 5095–5115. https://doi.org/10.1175/2008MWR2387.1

UPP. (2019). Universal post processor, version 4.1.0. [Software]. Retrieved from https://github.com/NOAA-EMC/UPP/commit/3b42c2a4999283bce18f7a45bccf4610703ad556

Weisman, M. L., Davis, C., Wang, W., Manning, K. W., & Klemp, J. B. (2008). Experiences with 0–36-h explicit convective forecasts with the WRF-ARW model. *Weather and Forecasting*, *23*(3), 407–437. https://doi.org/10.1175/2007WAF2007005.1

Wheatley, D. M., Yussouf, N., & Stensrud, D. J. (2014). Ensemble Kalman filter analyses and forecasts of a severe mesoscale convective system using different choices of microphysics schemes. *Monthly Weather Review*, *142*(9), 3243–3263. https://doi.org/10.1175/MWR-D-13-00260.1

Wolff, J. K., Harrold, M., Hertneky, T., Aligo, E., Carley, J. R., Ferrier, B., et al. (2016). Mesoscale Model Evaluation Testbed (MMET): A resource for transitioning NWP innovations from Research to Operations (R2O). *BAMS*, *97*(11), 2135–2147. https://doi.org/10.1175/BAMS-D-15-00001.1

Zhang, Y., Stensrud, D. J., & Clothiaux, E. E. (2021). Benefits of the Advanced Baseline Imager (ABI) for ensemble-based analysis and prediction of severe thunderstorms. *Monthly Weather Review*, *149*(2), 313–332. https://doi.org/10.1175/MWR-D-20-0254.1

Zhou, L. S.-J., Lin, J.-H., Chen, L. M., Harris, X., Chen, & Rees, S. L. (2019). Toward convective-scale prediction within the next generation global prediction system. *BAMS*, *100*(7), 1225–1243. https://doi.org/10.1175/BAMS-D-17-0246.1