JGR Atmospheres

Key Points:

- Thompson microphysics scheme has the most accurate upper-level simulated clouds, with National Severe Storms Laboratory producing too few
- Changing the planetary boundary layer scheme from Mellor-Yamanda-Nakanishi-Niino (MYNN) to Shin-Hong or Eddy-Diffusivity Mass-Flux resulted in slightly lower cloud accuracy
- RUC land surface model produces too many clouds compared to Noah, and is further enhanced when using the MYNN surface layer instead of Global Forecasting System

Correspondence to:
S. M. Griffin,
sarah.griffin@ssec.wisc.edu

Evaluating the Impact of Planetary Boundary Layer, Land Surface Model, and Microphysics Parameterization Schemes on Cold Cloud Objects in Simulated GOES-16 Brightness Temperatures

Sarah M. Griffin[1] , Jason A. Otkin[1] , Sharon E. Nebuda[1], Tara L. Jensen[2], Patrick S. Skinner[3,4,5], Eric Gilleland[2] , Timothy A. Supinie[6], and Ming Xue[6]

[1]Cooperative Institute for Meteorological Satellite Studies, University of Wisconsin-Madison, Madison, WI, USA, [2]National Center for Atmospheric Research/Research Applications Laboratory and Developmental Testbed Center, Boulder, CO, USA, [3]Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, OK, USA, [4]NOAA/National Severe Storms Laboratory, University of Oklahoma, Norman, OK, USA, [5]School of Meteorology, University of Oklahoma, Norman, OK, USA, [6]Center for Analysis and Prediction of Storms, University of Oklahoma, Norman, OK, USA

Abstract Infrared brightness temperatures (BTs) from the Geostationary Observing Environmental Satellite-16 Advanced Baseline Imager are used to examine the ability of several microphysics and planetary boundary layer (PBL) schemes, as well as land surface models (LSM) and surface layers, to simulate upper-level clouds. Six parameterization configurations were evaluated. Cloud objects are identified using the Method for Object-Based Diagnostic Evaluation (MODE) and analyzed using the object-based threat score, mean-error distance, and pixel-based metrics including the mean absolute error and mean bias error (MBE) for matched objects where the displacement between objects has been removed. Objects are identified using either a fixed BT threshold of 235 K or the 6.5th percentile of BTs for each model configuration. Analysis of the MODE-identified cloud objects shows that, compared to a configuration with the Thompson microphysics scheme, Mellor-Yamanda-Nakanishi-Niino (MYNN) PBL, Global Forecasting System (GFS) surface layer, and Noah LSM, the configuration employing the National Severe Storms Laboratory microphysics produced more cloud objects with higher BTs. Changing the PBL from MYNN to Shin-Hong or Eddy-Diffusivity Mass-Flux also resulted in a slightly lower accuracy, though these changes result in configurations which more accurately reproduced the number of observation cloud objects and slightly reduced the high MBE. Changing the LSM from Noah to RUC reduces forecast accuracy by producing too many cloud objects with too low BTs. As the forecast hour increases, this accuracy reduction increases at a greater rate than occurred when changing the microphysics or PBL scheme and is further enhanced when using the MYNN surface layer rather than the GFS.

Plain Language Summary Weather prediction models consist of many different parameters, which impact the resulting forecast. In this paper, we compared cloud forecasts generated using different model parameters to represent cloud processes, the bottom layer of the atmosphere, and the model surface. These model parameters are being considered for inclusion in future operational versions. It was found, the changes to the microphysics scheme dictating cloud processes had the largest impact on the accuracy of cloud forecasts, with the Thompson scheme producing more accurate clouds than the National Severe Storms Laboratory scheme. Cloud forecasts were more accurate when the bottom layer of the atmosphere was defined using the Mellor-Yamanda-Nakanishi-Niino (MYNN) scheme instead of Shin-Hong or an Eddy-Diffusivity Mass-Flux scheme. Changing the land surface model from Noah to RUC also reduced cloud forecast accuracy. This reduction in accuracy increases as the forecast hour increases, at a greater rate than occurred when changing the microphysics scheme or the scheme defining the bottom layer of the atmosphere. This reduction is further enhanced when using the MYNN surface layer rather than the Global Forecasting System.

1. Introduction

An accurate depiction of the spatial extent and temporal evolution of upper-tropospheric clouds is necessary to produce skillful convective weather forecasts. Convective storm clouds are difficult to predict, owing to complex nonlinear interactions between different cloud hydrometeor species and the impact of the local thermodynamic environment and mesoscale processes on their evolution (R. Johnson & Mapes, 2001). Through the inclusion of additional cloud processes and the ability to predict more than one moment of the particle size distribution, cloud microphysics schemes have become more realistic in recent years (e.g., Jones et al., 2018; Thompson et al., 2016). Improvements have also been made to other model components such as the planetary boundary layer (PBL) schemes (Y. Huang & Peng, 2017; Sušelj & Sood, 2010). These and other developments have contributed to more accurate cloud forecasts in convection-allowing models (CAMs) (e.g., Benjamin et al., 2016).

Despite the difficulty in predicting clouds and associated convective weather, these predictions have important socioeconomic impacts. For example, accurate cloud forecasts are important for determining when and where severe weather may occur (Cintineo et al., 2013; Mecikalski & Bedka, 2006; Purdom, 1993; Sieglaff et al., 2011). Thunderstorms account for most of the air traffic delays in the United States (Kaplan et al., 2005; Mecikalski et al., 2007; Murray, 2002), and the Federal Aviation Administration predicts that passenger volume on the US commercial carriers will increase almost 40% in the next 20 years (Federal Aviation Administration, 2018). In addition, cloud characteristics such as cloud-top height and optical depth also greatly impact solar energy prediction; thus, inaccurate cloud forecasts present a challenge to solar power integration (Haupt et al., 2018; Lee et al., 2017). These examples demonstrate the need to continue to improve the accuracy of cloud forecasts in numerical weather prediction (NWP) models.

The purpose of this paper is to contribute to ongoing efforts within the modeling community to increase the accuracy of cloud forecasts by assessing the ability of different parameterization schemes to produce accurate upper-level cloud objects primarily associated with deep convection. The schemes included in this analysis are being considered for inclusion in future operational versions of the Finite Volume Cubed Sphere (Lin, 2004; Putman & Lin, 2007) limited area model (FV3-LAM). This assessment uses output from a suite of experimental FV3-LAM simulations run by the Center for the Analysis and Prediction of Storms during the 2019 Hazardous Weather Testbed (HWT) Spring Forecasting Experiment (Clark et al., 2020; Gallo et al., 2017). The FV3-LAM simulations employed different microphysics and PBL parameterization schemes, as well as different land surface models (LSM) and surface layers.

One approach for assessing the accuracy of clouds in NWP models is to compare observed and simulated satellite infrared (IR) brightness temperatures (BTs), as IR BTs are highly sensitive to cloud-top properties and the horizontal distribution of clouds. Unlike visible imagery, IR BTs are available day and night, which makes them very valuable for model verification. Therefore, multiple studies have used observed BTs to assess the accuracy of model forecasts (e.g., Bikos et al., 2012; Cintineo et al., 2014; Feltz et al., 2009; Grasso & Greenwald, 2004; Grasso et al., 2008, 2014; Griffin et al., 2017a, 2017b; Jankov et al., 2011; H. Jin et al., 2014; Morcrette, 1991; Otkin & Greenwald, 2008; Otkin et al., 2009, 2017; Thompson et al., 2016). This analysis will conduct a detailed assessment of forecast accuracy using observed and simulated IR BTs from the Geostationary Observing Environmental Satellite (GOES)-16 Advanced Baseline Imager (ABI; Schmit et al., 2017) for forecasts run across the contiguous United States (CONUS).

Various techniques can be used to assess forecast accuracy, such as conventional grid point metrics, neighborhood methods, and object-based verification (Gilleland et al., 2009, 2010; Jolliffe & Stephenson, 2012). Object-based methods present a powerful way to analyze cloud cover forecast accuracy because traditional metrics typically penalize CAM forecasts for displacement errors between the forecast and observed cloud objects. Therefore, object-based methods and IR BTs have been used to assess forecast accuracy (Griffin et al., 2017a, 2017b, 2020; Jones et al., 2018, 2020; Senf et al., 2018; Skinner et al., 2016, 2018). In this analysis, cloud objects are identified in simulated and observed satellite imagery using the Method for Object-Based Diagnostic Evaluation (MODE; Bullock et al., 2016; Davis et al., 2006a, 2006b, 2009). MODE can be used to not only assess spatial displacement errors, but also to analyze other aspects of cloud objects such as their size. Many studies have used MODE to assess convective forecast skill, such as analyzing precipitation forecasts (Bytheway & Kummerow, 2015; Cai & Dumais, 2015; Wolff et al., 2014), cloud cover (Griffin

et al., 2017a, 2017b, 2020; Mittermaier & Bullock, 2013), or synoptic features (Mittermaier et al., 2016). In addition, this analysis will use the mean-error distance (MED) (Gilleland, 2017) to summarize the amount of displacement between features in the observation and forecast fields, as well as other properties such the amount of overlap.

This paper is organized as follows. Section 2 describes the data sets and methodology, with results presented in Section 3. This section includes complementary analyses that examine the forecast accuracy using object-based and pixel-based methods for cloud objects identified using fixed or variable BT thresholds for each model configuration. Finally, discussion and conclusions follow in Section 4.

2. Data

2.1. Model Configurations

This analysis uses output from a suite of six FV3-LAM configurations employing different parameterization schemes during the 2019 NOAA HWT Spring Forecasting Experiment. All forecasts are initialized at 00 UTC each weekday during May 14–31, 2019 and then run for 60 h. Their long duration allows us to examine changes in forecast accuracy during two diurnal cycles. Forecast hours (FHs) 0–5 are not included in this analysis to reduce the impact of model spin-up processes on the forecast cloud field because the model starts from a cloud-free state.

All forecasts are run on a domain covering the CONUS at 3 km grid spacing with 64 vertical levels, including 13 levels between 400 and 100 hPa. The model domain and a representative example of the simulated BTs are shown in Figure 1, with a description of each model configuration found in Table 1. The Control configuration employs the Thompson microphysics scheme with the aerosol-aware capability turned off (Thompson et al., 2004, 2008), the Mellor-Yamanda-Nakanishi-Niino (MYNN) v3.6 PBL (Nakanishi & Niino, 2004, 2009), the Global Forecasting System (GFS) surface layer, the Noah LSM (Mitchell, 2005; Niu, 2011), and the North American Model (NAM) initial and boundary conditions. This configuration is referred to as the Control because it is the baseline upon which the other model configurations vary. One configuration explores the sensitivity to the cloud microphysics scheme, the National Severe Storms Laboratory (Mansell et al., 2010) scheme, hereafter referred to as MP-NSSL. A second set of experiments replaced the PBL scheme with the Shin-Hong (Shin & Hong, 2013) or hybrid eddy-diffusivity mass-flux (EDMF; Han et al., 2016) schemes, hereafter referred to as PBL-SH and PBL-EDMF, respectively. A final set of experiments employed the RUC v3.6+ (Smirnova et al., 2016) LSM with either the GFS or MYNN surface layers. These model configurations are denoted as LSM-RUC_SFC-GFS and LSM-RUC_SFC-MYNN. All of the simulations employ the Rapid Radiative Transfer Model for General Circulation Models (Clough et al., 2005) radiation scheme when computing radiative fluxes at the surface and in the atmosphere, with in-cloud fluxes directly coupled to the cloud properties via the prognostic mixing ratios and diagnosed effective radii for each cloud species (Thompson et al., 2016).

2.2. Infrared Brightness Temperatures

The verification data set is observed BTs from the GOES-16 ABI sensor. GOES-16 is located over the equator at 74°W and has 16 bands in the visible, shortwave, and longwave IR. This analysis uses the 10.3 μm BTs, which have a 2-km pixel spacing at nadir. The observed BTs are calculated by remapping the observed radiances to the model grid using an area-weighted average of all the observed pixels overlapping a given grid box and then converting them to BTs. Simulated ABI BTs were generated using the Community Radiative Transfer Model (CRTM V2+) and vertical profiles of temperature, specific humidity, and cloud hydrometeors (Ding et al., 2011; Han et al., 2005). For cloudy grid points, a diagnosed effective radius for each microphysical species (cloud water, rainwater, ice, snow, and graupel) that is consistent with the assumptions made by the parameterization scheme (Otkin et al., 2007) is input to the CRTM. The National Polar-Orbiting Operational Environmental Satellite System IR emissivity database is used to model land surface emissivity (Han et al., 2005).
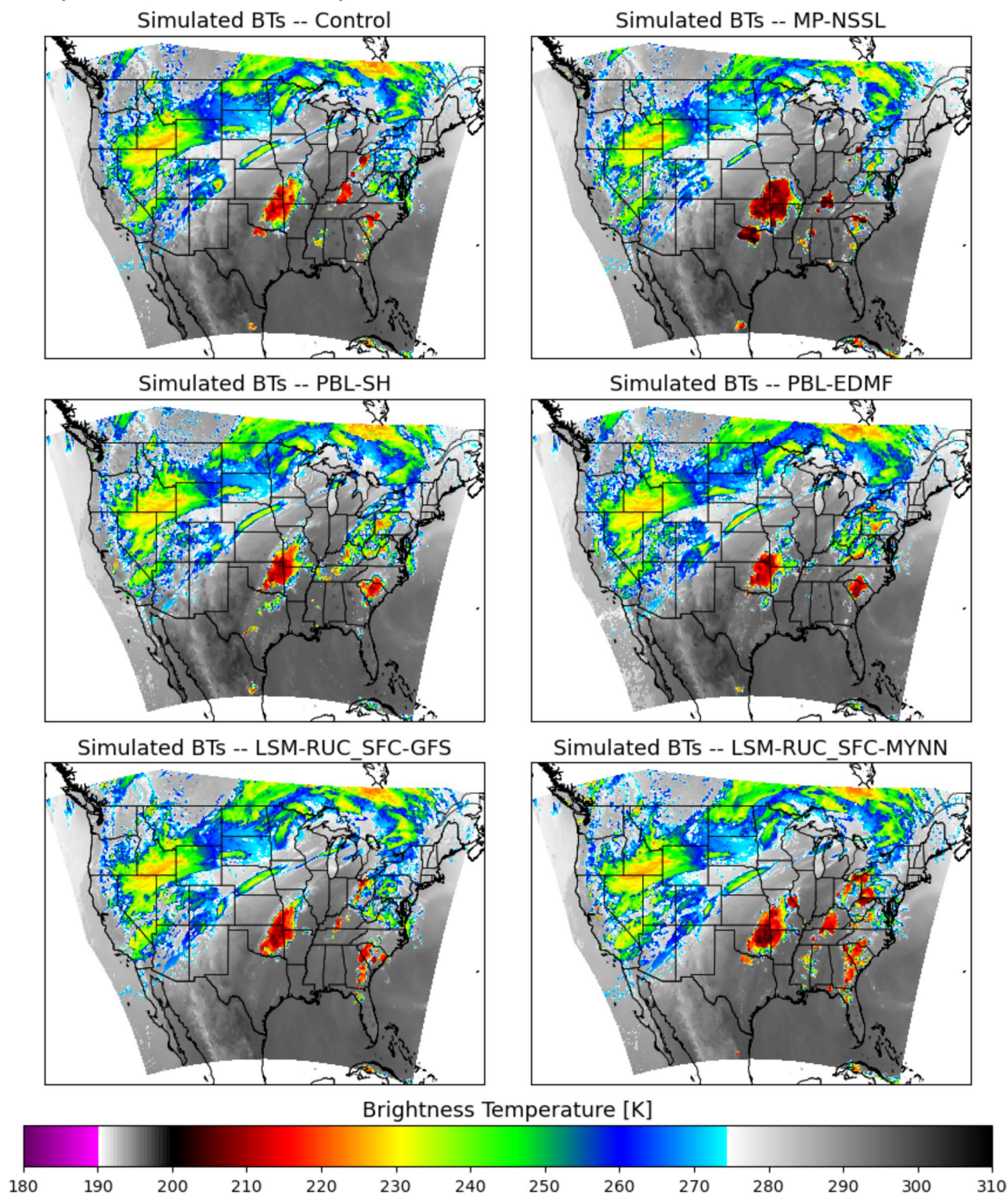
Figure 1. An example of simulated brightness temperatures from 20190522 at 00 UTC valid on 20190523 at 00 UTC for the six different Finite Volume Cubed Sphere-limited area model model configurations used in this analysis. A description of the configurations can be found in Table 1. BTs, brightness temperatures; EDMF, hybrid eddy-diffusivity mass-flux; GFS, Global Forecasting System; LSM, land surface models; MYNN, Mellor-Yamanda-Nakanishi-Niino; PBL, planetary boundary layer; SH, Shin-Hong.

Table 1
Description of Experiments Used in This Analysis

Name	Microphysics scheme	Planetary boundary layer scheme	Surface layer	Land surface model
Control	Thompson	MYNN	GFS	Noah
MP-NSSL	National Severe Storms Laboratory	MYNN	GFS	Noah
PBL-SH	Thompson	Shin-Hong	GFS	Noah
PBL-EDMF	Thompson	EDMF	GFS	Noah
LSM-RUC_SFC-GFS	Thompson	MYNN	GFS	RUC
LSM-RUC_SFC-MYNN	Thompson	MYNN	MYNN	RUC

Abbreviations: EDMF, hybrid eddy-diffusivity mass-flux; GFS, Global Forecasting System; LSM, land surface models; MYNN, Mellor-Yamanda-Nakanishi-Niino; PBL, planetary boundary layer; SH, Shin-Hong.

3. Methodology

Cold cloud objects are identified in the observed and simulated 10.3 μm BT imagery using MODE (Davis et al., 2006a, 2006b, 2009). MODE uses a process called convolutional thresholding to identify objects. An example of identifying GOES observation objects can be seen in Figure 2. In this analysis, BTs are smoothed using a convolution radius of 5 grid points, which is 15 km (Figure 2b). This radius allows for the analysis of large and small-scale objects, as Cai and Dumais (2015) state a range of 2–8 grid points is necessary to identify convective storm objects in ∼4-km resolution radar imagery. From these smoothed BTs, objects are identified (Figure 2c) and then various object attributes, such as location and size, are computed for each object. Finally, the interest score is calculated, which equates the similarity of matching observation and forecast objects (Bullock et al., 2016; Jensen et al., 2020). The interest score is a weighted combination of the object pair attributes and ranges in value from 0 to 1, with 1 being a perfect match. The attributes and user-defined weights applied in this study (Table 2) are the same as those employed by Griffin et al. (2017a, 2017b, 2020). It should be noted that MODE uses a centroid distance weight that is the user-defined weight multiplied by the ratio of the objects' areas. Overall, this analysis prioritizes distance and size comparisons between the objects. More emphasis is placed on the displacement between the centroids of the objects and the ratio of the objects' areas because both the boundary distance and ratio of the intersection area can be artificially high, especially when large objects only slightly overlap each other, or a larger object fully encloses a smaller object.

Interest scores are also used to identify clusters of objects. Clusters are defined as one or more observation objects matched with one or more forecast objects (Bullock et al., 2016; Jensen et al., 2020), with a minimum interest score of 0.65. For example, a large forecast object can be matched with multiple smaller observation objects, therefore allowing these observation objects to be considered part of a matched pair. In this case, the attributes of the observation cluster, including its area and distance between the matched clusters, are used instead of the individual objects' attributes. If clusters are not used, smaller objects might not have a match. These objects (individually or as part of a cluster) are used to assess the model forecast accuracy using the methods below.

3.1. Object-Based Threat Score

The accuracy of the forecast cloud objects is assessed using the object-based threat score (OTS; A. Johnson & Wang, 2013). The OTS is calculated using object sizes and interest score between matching pairs of objects based on the equation:

$$OTS = \frac{1}{A_f + A_o}\left[\sum_{p=1}^{P} I^p\left(a_f^p + a_o^p\right)\right], \quad (1)$$

where A_f and A_o represent the total spatial area of all matched and unmatched forecast and observation objects, respectively. P represents the number of matched forecast and observation object pairs, with I^p
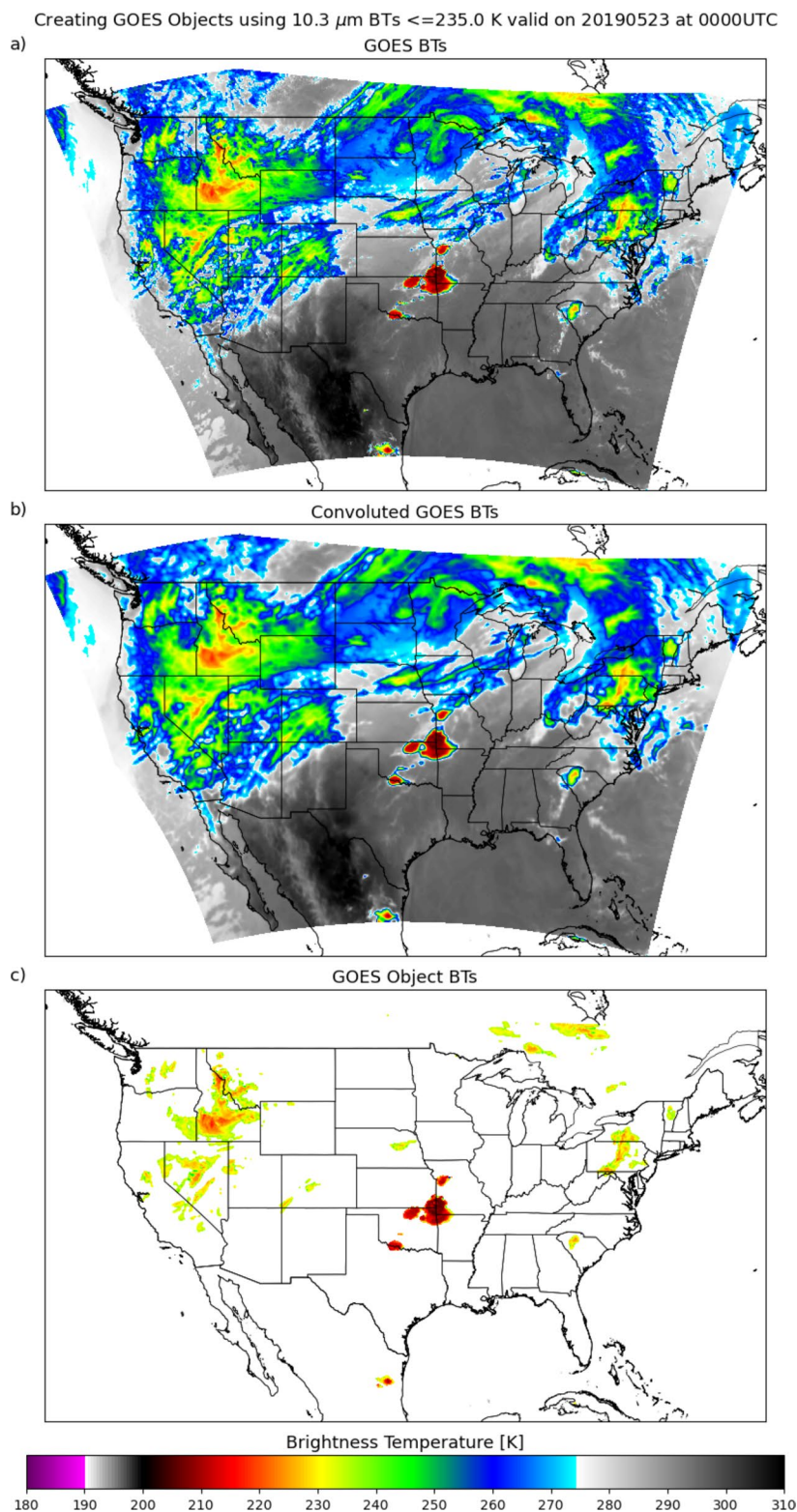
Figure 2. Example of the convolution thresholding process to identify Geostationary Observing Environmental Satellite (GOES) observation objects on 20190523 at 00 UTC. (a) GOES observed brightness temperatures (BTs). (b) Smoothed GOES observed BTs based on a convolution radius of 5 grid points. (c) GOES observed BTs of identified objects from the smoothed BTs in Figure 2b.

Table 2
User-Defined Weights and Brief Description of the Object Pair Attributes Used in This Analysis

Object pair attribute	User-defined weight (%)	Description
centroid_dist	4 (25.0)	Distance between objects' "center of mass"
boundary_dist	3 (18.75)	Minimum distance between the objects
convex_hull_dist	1 (6.25)	Minimum distance between the polygons surrounding the objects
angle_diff	1 (6.25)	Orientation angle difference
area_ratio	4 (25.0)	Ratio of the forecast and observation objects' areas (or its reciprocal, whichever yields a lower value)
int_area_ratio	3 (18.75)	Ratio of the objects' intersection area to the lesser of the observation or forecast area (whichever yields a lower value)

representing the interest score between the matched forecast and observation object/cluster pair and a_f^p and a_o^p representing the areas of the forecast and observation objects/clusters in the matched pair, respectively. Since the OTS is based on a one-to-one correspondence, objects can only be used once in the calculation and therefore it is possible that not all objects will be included in a matched pair. Once an object has been used in the OTS calculation, either as an individual object or as part of a cluster, any additional matched pairs that include that object are no longer used in the summation over all of the object pairs. Thus, object and cluster matches are sorted from the highest to the lowest interest scores to calculate the highest possible OTS. Interest values are used until no more matching objects or clusters remain. The OTS has a scale from 0 to 1, with one representing a perfect forecast.

3.2. Mean-Error Distance

Another method used to quantify the distance between areas of interest in simulated and observed imagery, which in this analysis are MODE objects, is the MED. Unlike the centroid distance used in the MODE interest score calculation, which calculates the distance between the objects' centers of mass, MED calculates the shortest distance between every grid point identified as a forecast (observation) object to the closest grid point identified as an observation (forecast) object (Gilleland, 2017). MED from the observation object (O) to the forecast objects (F) is calculated using:

$$\mathrm{MED}\left(\mathrm{F},\mathrm{O}\right) = \frac{1}{\mathrm{N_O}} \sum_{x \epsilon O} \mathrm{d}\left(x,\mathrm{F}\right) \qquad (2)$$

where N_O is the number of grid points in the observation objects and d(x, F) is the shortest distance from a grid point in an observation object to a forecast object. Likewise, MED from a grid point in a forecast object to an observation object, is calculated using:

$$\mathrm{MED}\left(\mathrm{O},\mathrm{F}\right) = \frac{1}{\mathrm{N_F}} \sum_{x \epsilon F} \mathrm{d}\left(x,\mathrm{O}\right) \qquad (3)$$

with N_F being the number of grid points in the forecast objects and d(x, O) is the shortest distance from a forecast object grid point and an observation object. As a result, MED(F, O) $\neq$ MED(O, F). Both MED calculations have a perfect value of zero, which occurs when all grid points in the observation and forecast objects exactly overlap.

3.3. Pixel-Based Metrics

Two pixel-based metrics are used to assess the accuracy of the simulated BTs within the cold cloud objects. Unlike the object-based analysis described above, the mean absolute error (MAE) and mean bias error (MBE) will be calculated only for observation objects where all six of the model forecasts at a given time have a matched forecast-observation object pair. If a model configuration has a forecast object that is not matched to an observation object, then that forecast object is not used in the calculations. Therefore, these metrics are assessing the BTs within the cloud objects rather than the "goodness of the match" as is done
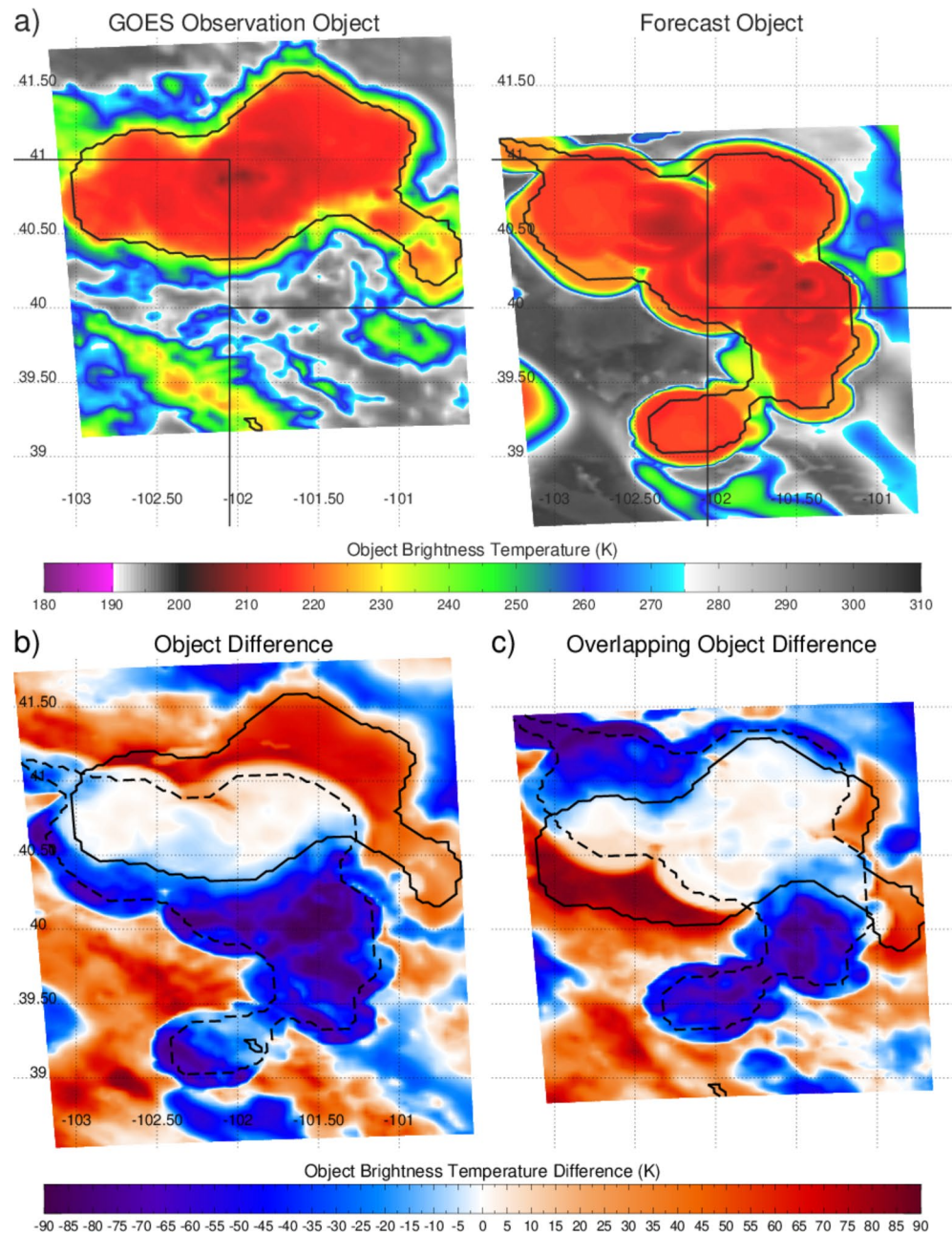
Figure 3. Example of removing displacement using the center latitude and longitude of Method for Object-Based Diagnostic Evaluation images. (a) Two matching object that are displaced. Difference between brightness temperatures when the objects (b) are not and (c) are overlapped. From Griffin et al. (2020). GOES, Geostationary Observing Environmental Satellite.

with the OTS and MED. The impact of spatial displacement errors between matched objects is removed by centering the objects using the object centroid latitude and longitude identified by MODE. An example of this centering can be seen in Figure 3. Since forecast and observation objects are not necessarily the same shape, these metrics are calculated over a box including grid points surrounding the identified cloud objects. The extent of this box, in each cardinal direction from the now overlapping objects' centers, is the largest distance that either the observation or forecast object extends from its respective center.

4. Results

4.1. Object-Based Threat Score

The OTS for cloud objects identified by MODE using a 235 K BT threshold (hereafter referred to as 235 K objects) is shown in Figure 4. The cloud object BT threshold of 235 K is used as it represents approximately the lowest 6.5% of BTs in the GOES-16 observations, which is similar to the climatological percentage of convective clouds (Chang & Li, 2005). Since the OTS uses the object areas, objects intersecting the edge of the domain are excluded from this analysis because the size of the entire object in that situation will be unknown. In addition, objects that are matched with an object intersecting the edge of the domain are excluded from this analysis. The OTS is also calculated for different sectors of the CONUS region, as seen in Figure 4a. An observation object is considered part of a given sector if at least 50% of the object area is located within that sector. Forecast objects matching observation objects in the sector are also considered to be in that sector. For unmatched forecast objects, the area of the object in the sector is considered in the OTs calculation.

Based on the sector analysis at the top of Figures 4a and 4c, all model configurations have the highest (lowest) OTS over the central (eastern) CONUS over the full forecast period. Also, all configurations have a less accurate OTS compared to Control in all sectors except for MP-NSSL over western CONUS. This is interesting because, unlike the Thompson scheme used in the Control, the NSSL microphysics scheme includes the graupel particle density and the mass and number of hail particles, and western CONUS has the lowest climatological probability of hail (Cintineo et al., 2012). LSM-RUC_SFC-GFS and LSM-RUC_SFC-MYNN have the lowest OTS compared to Control in all sectors, therefore using the RUC LSM leads to less accurate forecasts of cloud objects.

Over the full domain, the most accurate forecast varies based on FH. As seen in Figure 4b, between FHs 6 and 24, PBL-SH generally has the highest OTS. Therefore, the Shin-Hong PBL scheme generates more accurate cloud features than the MYNN for early FHs. LSM-RUC_SFC-GFS is also more likely to have a higher OTS than Control for the first 24 FHs. Control tends to be much more accurate during the latter part of the forecast period (Figure 4d), starting around FH 34. The LSM-RUC_SFC-MYNN forecasts have the steepest reduction in OTS as the FH increases, followed by LSM_RUC_SFC-GFS, which indicates that the rapid decrease in accuracy is due to the RUC LSM. For all of the model configurations, OTS decreases with increasing FH due to the limited predictability of convection at longer lead times. However, there is some periodicity in the OTS time series, as all configurations have a distinct local maximum between FH 22 and FH 28, followed by a second more diffuse maximum centered on FH 48.

To further investigate the underlying behavior of the OTS, box-and-whisker plots and Taylor diagrams of the number and area of observed and forecast objects, respectively, are shown in Figures 5 and 6. These results aggregate data from all of the forecasts over the full domain and are plotted as a function of FH. Taylor diagrams (Taylor, 2001) displaying the Pearson correlation coefficient along the curved y-axis, standard deviation along the x-axis, and root-mean-square difference (RMSD) along the dashed semi-circles are also shown. The RMSD is calculated using:

$$\text{RMSD} = \sqrt[2]{\frac{1}{N}\sum_{n=1}^{N}\left[\left(f_n - \overline{f}\right) - \left(o_n - \overline{o}\right)\right]^2} \qquad (4)$$

where f and o represent the number or area of the forecast and observation objects, respectively. Therefore, RMSD is the mean-squared error after accounting for biases (Wilks, 2011). A model configuration more accurately represents the observations if its location on the Taylor diagram is closer to the observation's location. For the Taylor diagram, the median number or area of objects from the box-and-whisker plots for FHs 6–60 are used for the calculations.

Based on Figure 5a, all model configurations overpredict the number of 235 K objects compared to the observations, with LSM-RUC_SFC-MYNN characterized by the highest number of cloud objects. This over-prediction in the number of cloud objects increases as the forecast progresses, with many of the configurations having more cloud objects during the second day. Since all of the configurations are initialized with a cold start, the lack of convective coverage during the first few FHs could lead to greater potential instability and an increase in convective clouds in subsequent forecasts (Wong & Skamarock, 2017), with
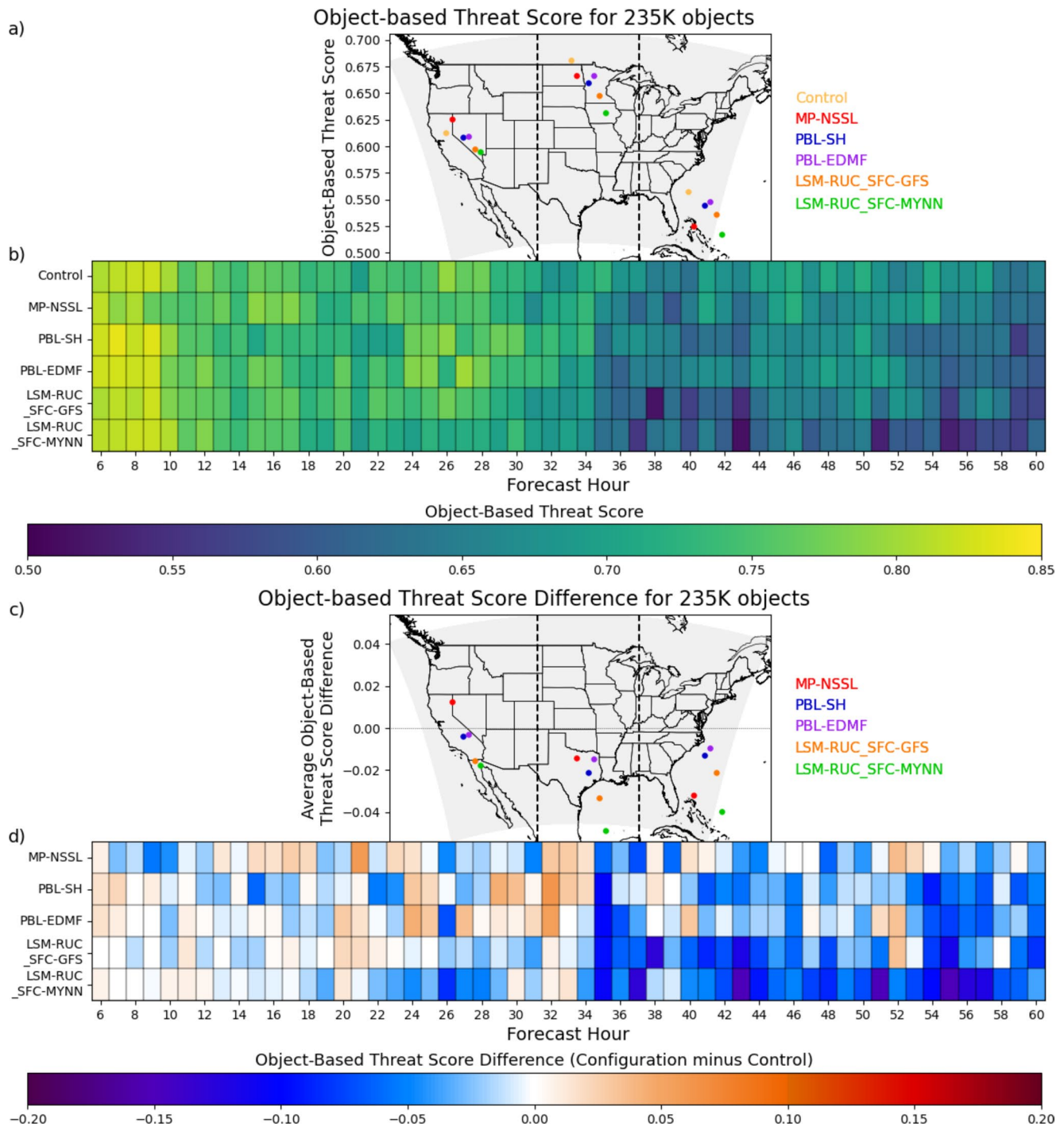
Figure 4. (a) Average object-based threat score (OTS) for objects identified using a threshold of 235 K over the full forecast period for each sector of the Contiguous United States. A perfect OTS has a value of 1. (b) OTS over the full domain by forecast hour (FH). (c) The difference between the configuration and Control OTS for each sector. (d) The difference between the configuration and Control OTS over the full domain by FH. Blue (red) indicates the configuration OTS is higher (lower) than Control. EDMF, hybrid eddy-diffusivity mass-flux; GFS, Global Forecasting System; LSM, land surface models; MYNN, Mellor-Yamanda-Nakanishi-Niino; PBL, planetary boundary layer; SH, Shin-Hong.
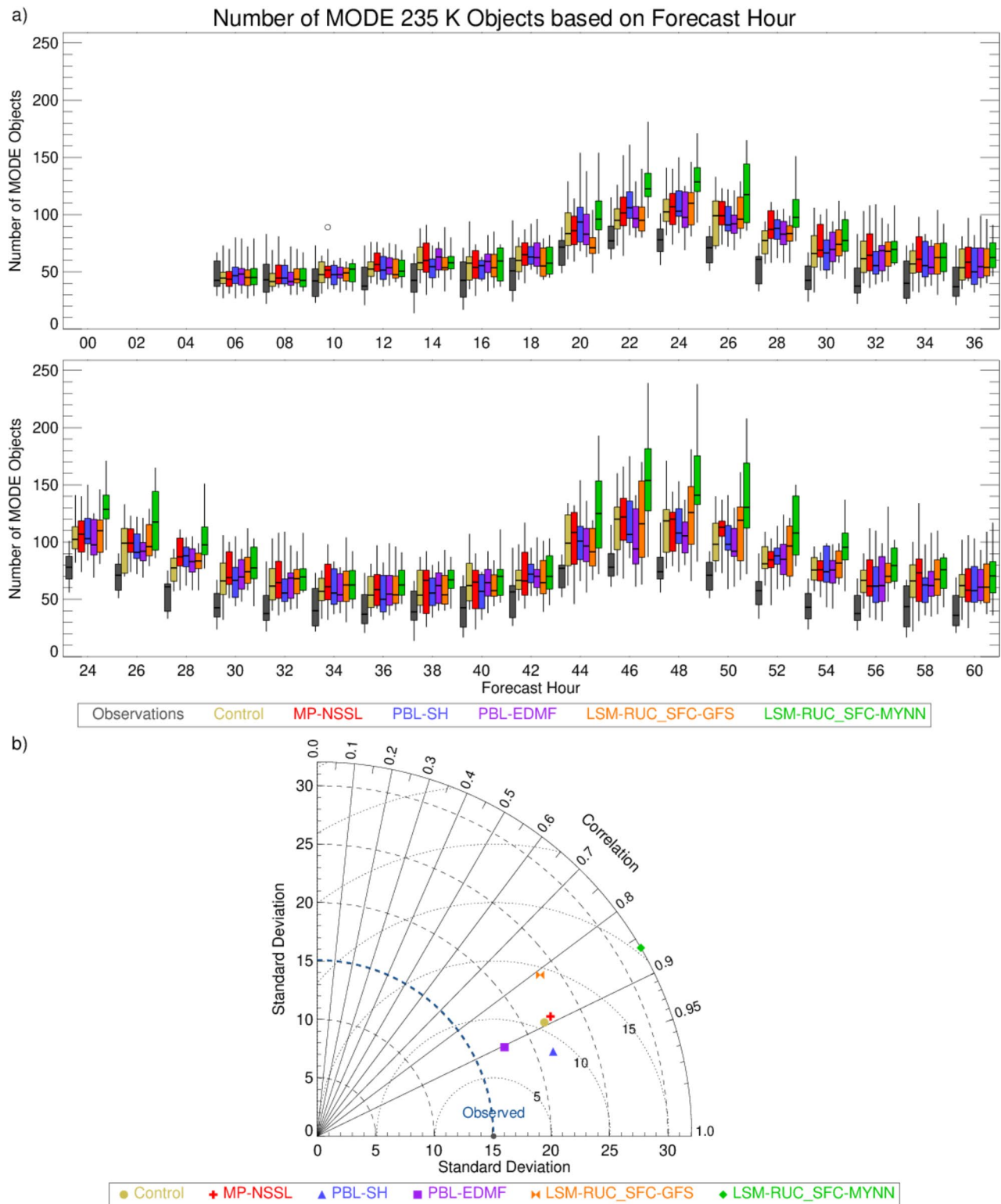
Figure 5. (a) Number of 235 K objects for the Geostationary Observing Environmental Satellite-16 observations and Finite Volume Cubed Sphere-limited area model model configurations as a function of forecast hour (FH) aggregated over all initializations. FHs 24–36 are repeated in the middle panel to facilitate intercomparison between the day 1 and 2 diurnal maxima. (b) Taylor diagram displays the Pearson correlation coefficient along the curved *y*-axis (solid lines), the root-mean-square difference along the dotted semi-circles in the plot, and standard deviation along the *x*-axis (dashed lines) for the number of 235 K objects. EDMF, hybrid eddy-diffusivity mass-flux; GFS, Global Forecasting System; LSM, land surface models; MODE, Method for Object-Based Diagnostic Evaluation; MYNN, Mellor-Yamanda-Nakanishi-Niino; PBL, planetary boundary layer; SH, Shin-Hong.
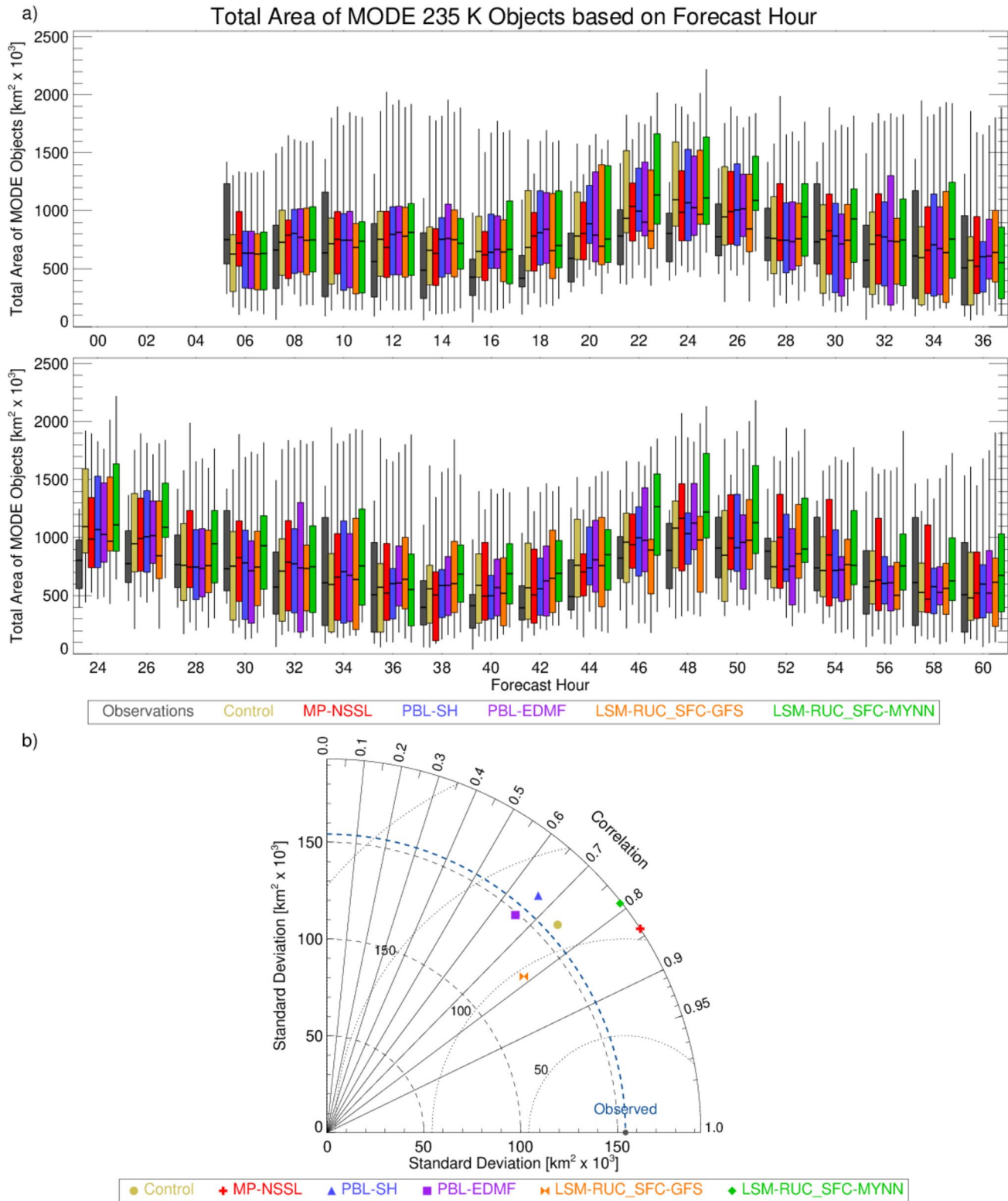
Figure 6. Same as Figure 5 but for area encompassed by 235 K objects. EDMF, hybrid eddy-diffusivity mass-flux; GFS, Global Forecasting System; LSM, land surface models; MODE, Method for Object-Based Diagnostic Evaluation; MYNN, Mellor-Yamanda-Nakanishi-Niino; PBL, planetary boundary layer; SH, Shin-Hong.

this overall increase in the number of objects corresponding to the decrease in the OTS (Figure 4). However, a local maximum in the number of objects is observed corresponding to the local maximum in OTS. Inspection of the critical success index and equitable threat scores for these times (not shown) indicates that the more numerous cloud objects do not increase the OTS by increasing the probability of observed and forecast cloud objects overlapping or the likelihood of a random forecast being accurate.

The Taylor diagram in Figure 5b shows that the spread in the median number of forecast objects is larger than occurred in the observations. The PBL-SH and PBL-EDMF configurations have the lowest RMSD, which indicates that changing the PBL scheme from the MYNN in the Control to either the Shin-Hong or EDMF reduces the median number of 235 K objects after accounting for the high bias in the number of objects (Figure 5a). The PBL-SH configuration exhibits a stronger correlation to the observations than the PBL-EDMF. The LSM-RUC_SFC-MYNN configuration has the largest standard deviation and RMSD because of the increase in the number of objects at later FHs, which is consistent with the decrease in OTS shown in Figure 4. Comparison of the LSM-RUC_SFC-GFS and LSM-RUC_SFC-MYNN results show that use of the GFS surface layer rather than the MYNN led to more accurate cloud forecasts based on the smaller standard deviation and RMSD; however, this configuration also had a slightly lower correlation.

The overall area of the cloud objects can be seen in Figure 6a. Increases in the area encompassed by the forecast and observation objects around FHs 24 and 48 correspond to the local maxima in OTS at those times (Figure 4), and the average size of forecast and observation objects at these times decreases (not shown). There is a slight increase in the area of forecast objects on the second day, however, the average area of each object decreases given the increase in the number of objects (Figure 5a). The Taylor Diagram (Figure 6b) shows that LSM-RUC_SFC-GFS has the lowest spread and RMSD overall. This indicates that the RUC LSM leads to smaller errors in the area encompassed by objects, once accounting for the overall bias in the area of objects, than occurred when using the Noah LSM. However, it should also be noted that the Control configuration has the third lowest RMSD and better identifies the spread in the object area than forecasts using the LSM-RUC_SFC-GFS configuration.

The impact of the area encompassed by objects on the OTS can be observed by breaking the OTS into its three main parts as seen in Figure 7: the percent of observation objects matched (a_o/A_o), the percent of forecast objects matched (a_f/A_f), and the average interest score between matched objects $\left(\frac{1}{P} \sum\limits_{p=1}^{P} I^p \right)$. PBL-SH is overall more accurate than Control for FHs 6–24 because more of the observation area is matched to PBL-SH forecast objects. For LSM-RUC_SFC-GFS, more forecast object area is matched for FHs 6–24 compared to Control forecast object area. The periodicity in the OTS is caused by local maxima in the average interest scores due to local maxima in the centroid distance and boundary layer distance attribute scores (not shown).

4.2. Mean-Error Distance

The MED statistics for the same objects used to calculate the OTS in Figure 4 are shown in Figure 8. Overall, the MED from the forecast objects to the observation objects [MED(O, F)] (Figures 8a and 8b) is larger than the MED from the observation objects to the forecast objects [MED(F, O)] (Figures 8c and 8d). This behavior is generally a result of an increased number of grid points being identified as parts of cloud objects in the forecasts compared to the observations. This can especially be seen in the central and eastern CONUS, where there are more forecast grid points than object grid points, compared to the western CONUS, where the number of observation object grid points is close to the number of forecast grid points (Figure 8e) and the MEDs are more similar. Increasing the number of grid points could result in more overlapping grid points, which have an error distance of zero, if the additional object grid points increase the object's size enough to overcome the displacement between objects. Control has the lowest MED from the forecast to the observation object over the full domain and in most sectors. The MED(O, F) for MP-NSSL, PBL-SH, and PBL-EDMF is ~1.5 grid points worse than the Control, which is an increase of 5%. Therefore, the PBL and microphysics scheme changes have only a slight impact on MED(O, F). Looking at the number of grid points defined as being part of cloud objects (Figure 8f), LSM-RUC_SFC-GFS has the lowest average number of forecast object grid points and the second highest MED(F, O). However, MP-NSSL has the highest
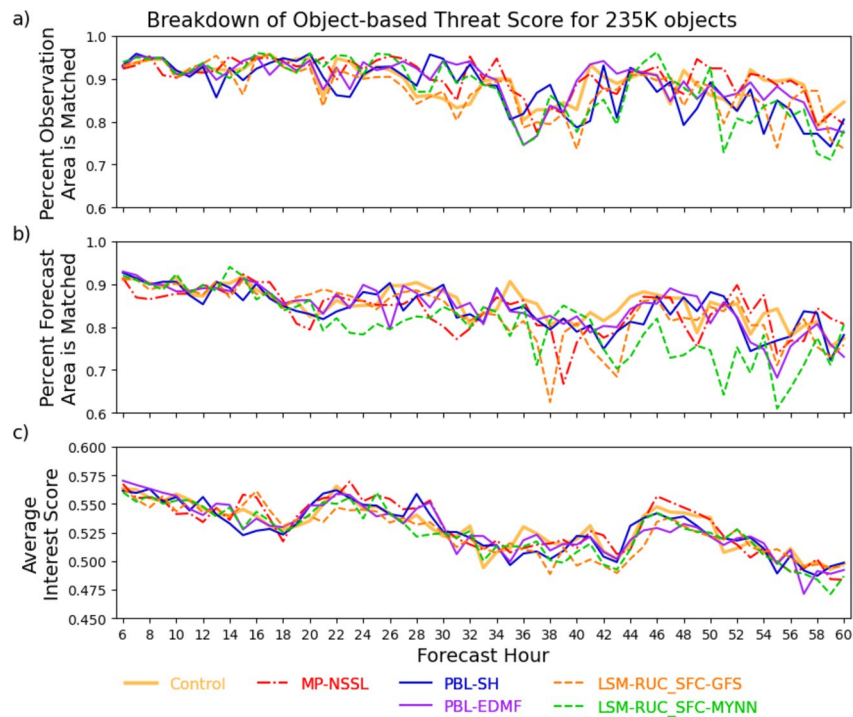
Figure 7. Breakdown of the object-based threat score for objects using a threshold of 235 K. (a) Percent of 235 K observation object area that is matched to a forecast object, (b) percent of 235 K forecast object area that is matched to an observation object, and (c) average interest score for all six Finite Volume Cubed Sphere-limited area model configurations by forecast hour. EDMF, hybrid eddy-diffusivity mass-flux; GFS, Global Forecasting System; LSM, land surface models; MYNN, Mellor-Yamanda-Nakanishi-Niino; PBL, planetary boundary layer; SH, Shin-Hong.

MED(F, O), which indicates that the presence of more forecast object grid points is not solely responsible for a lower MED(F, O). This is true in each of the sectors and over the full domain. LSM-RUC_SFC-MYNN has the highest number of forecast object grid points, and this difference is most evident in central CONUS.

The local minimum in MED(O, F) occurs at the same time as the local maximum in OTS (Figure 4). There is also a corresponding local maximum in the number of grid points identified as being part of cloud objects in both the observations and forecasts. Therefore, the decrease in MED(O, F) at these times is potentially caused by more forecast and observation object grid points overlapping. This behavior is especially evident for the local minimum in MED(O, F) around FH 48, where a local maximum in the percent of forecast object grid points overlapping with observation object grid points is seen (Figure 9). However, the local minimum around FH 22 does not correspond to an increase in overlapping grid points. Thus, the forecast objects are closer to the observation objects, however that does not necessarily result in additional overlap.

4.3. Pixel-Based Metrics

In this section, pixel-based metrics are utilized to analyze the observed and simulated BTs associated with the matched object pairs. To allow for a direct comparison of forecast object BTs, unlike the analyses shown in Sections 4.1 and 4.2, only observation objects which have a match in all of the configurations are included in this analysis. This approach was chosen so that the MAE and MBE will not be skewed by additional objects produced by any model configuration; however, it does mean that some objects in the previous analysis will be excluded.

Time series depicting the evolution of the MAE and MBE associated with these matched object pairs are shown in Figure 10, with the average MAE and MBE for all forecasts shown in Table 3. On average, Control has the lowest MAE over the full domain (Figure 10b) and for each sector (Figure 10a), indicating that it
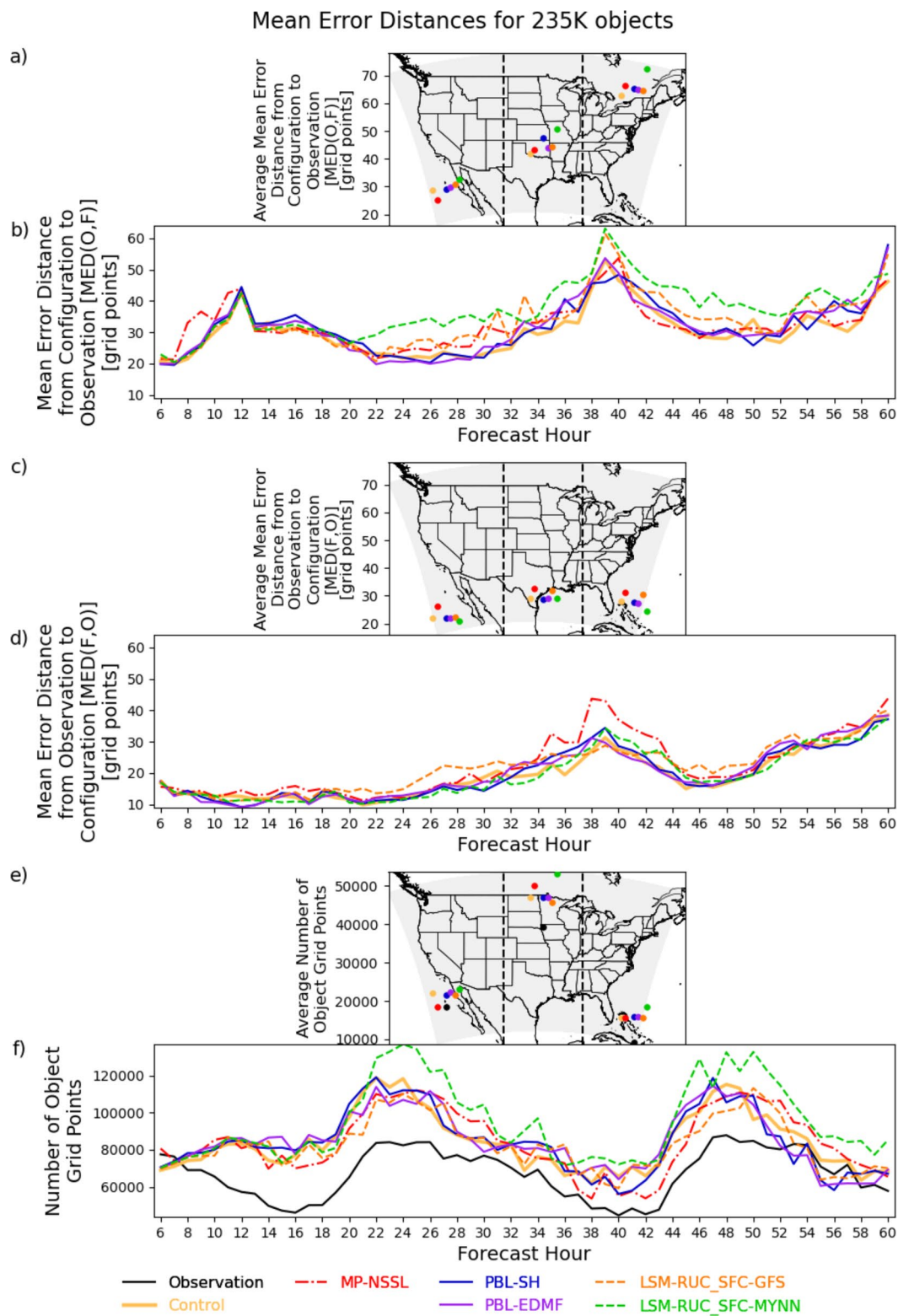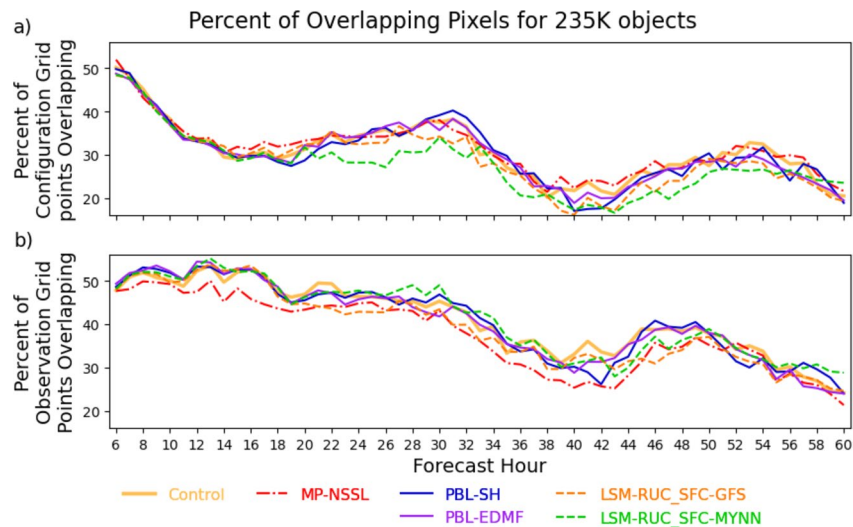
Figure 8

Figure 9. (a) Percent of grid points identified as a Method for Object-Based Diagnostic Evaluation (MODE) object for each configuration that overlap with a MODE observation object. (b) Percent of grid points identified as a MODE observation object that overlap a MODE forecast object for each model configuration. EDMF, hybrid eddy-diffusivity mass-flux; GFS, Global Forecasting System; LSM, land surface models; MYNN, Mellor-Yamanda-Nakanishi-Niino; PBL, planetary boundary layer; SH, Shin-Hong.

has the most accurate BTs in the matched cloud objects. Similar errors occurred during the PBL-EDMF, PBL-SH, and LSM-RUC_SFC-GFS simulations; thus, changing the PBL scheme or the land surface model has minimal impact on the BTs within the cold cloud objects. The average MAE for MP-NSSL is higher than Control, indicating that the Thompson microphysics scheme produces more accurate BTs in the forecast cloud objects than the NSSL MP scheme.

Inspection of the bias time series (Figure 10d) shows that the Control configuration has a slightly positive MBE, indicating that the simulated BTs for matched object pairs are higher than observed. This high bias is the largest in the western CONUS sector (Figure 10c). The PBL-EDMF, LSM-RUC_SFC-GFS, and PBL-SH configurations have slightly lower MBEs than Control. Therefore, using the EDMF PBL scheme instead of MYNN or RUC LSM instead of Noah leads to lower BTs. For reference, the accuracy of GOES-16 10.3 μm BTs is 0.025 K (Wu & Schmit, 2019), so therefore these small MBEs are still significant. While the average biases over the full domain are also closer to zero, the BTs are still slightly less accurate compared to the Control based on the MAE (Figure 10a). MP-NSSL has the largest MBE, indicating that it has higher BTs compared to the Control configuration employing the Thompson scheme. Overall, the MBE is correlated (Pearson correlation coefficient of -0.57) with the difference between the size of the forecast object and observation object, as a greater difference resulting in a lower MBE.

Also of note from Figure 10 is the local maxima in MAE around FHs 24 and 48 that roughly corresponds to the local maxima in OTS at those times (Figure 4). This result is counterintuitive because these metrics are showing that the model configurations are both more or less accurate at the same time. However, it is important to reiterate that the MAE and OTS analyze different object features. The MAE compares the BTs between matched object pairs, whereas the OTS compares the size and location of matched object pairs in addition to accounting for unmatched objects. Therefore, if a matched object pair has the exact same size and location, but the BTs differ within them, the object pair will have the highest possible interest score of 1 and thus the largest OTS while also having an MAE that is not the lowest possible value of 0 K. A deeper

Figure 8. Mean-error distance (MED) for objects identified using a threshold of 235 K by forecast hour (FH). (a) Average MED from configuration to observation [MED(O, F)] over the full forecast period for each sector on the contiguous United States. (b) MED(O, F) over the full domain by FH. (c) Same as Figure 8a but for MED from observation to configuration [MED(F, O)]. (d) Same as Figure 8b but for MED(F, O). (e) The average number of grid points used in the MED calculation for each model configuration for each sector over the full forecast period. (f) The average number of grid points used in the MED calculation for each model configuration. EDMF, hybrid eddy-diffusivity mass-flux; GFS, Global Forecasting System; LSM, land surface models; MYNN, Mellor-Yamanda-Nakanishi-Niino; PBL, planetary boundary layer; SH, Shin-Hong.
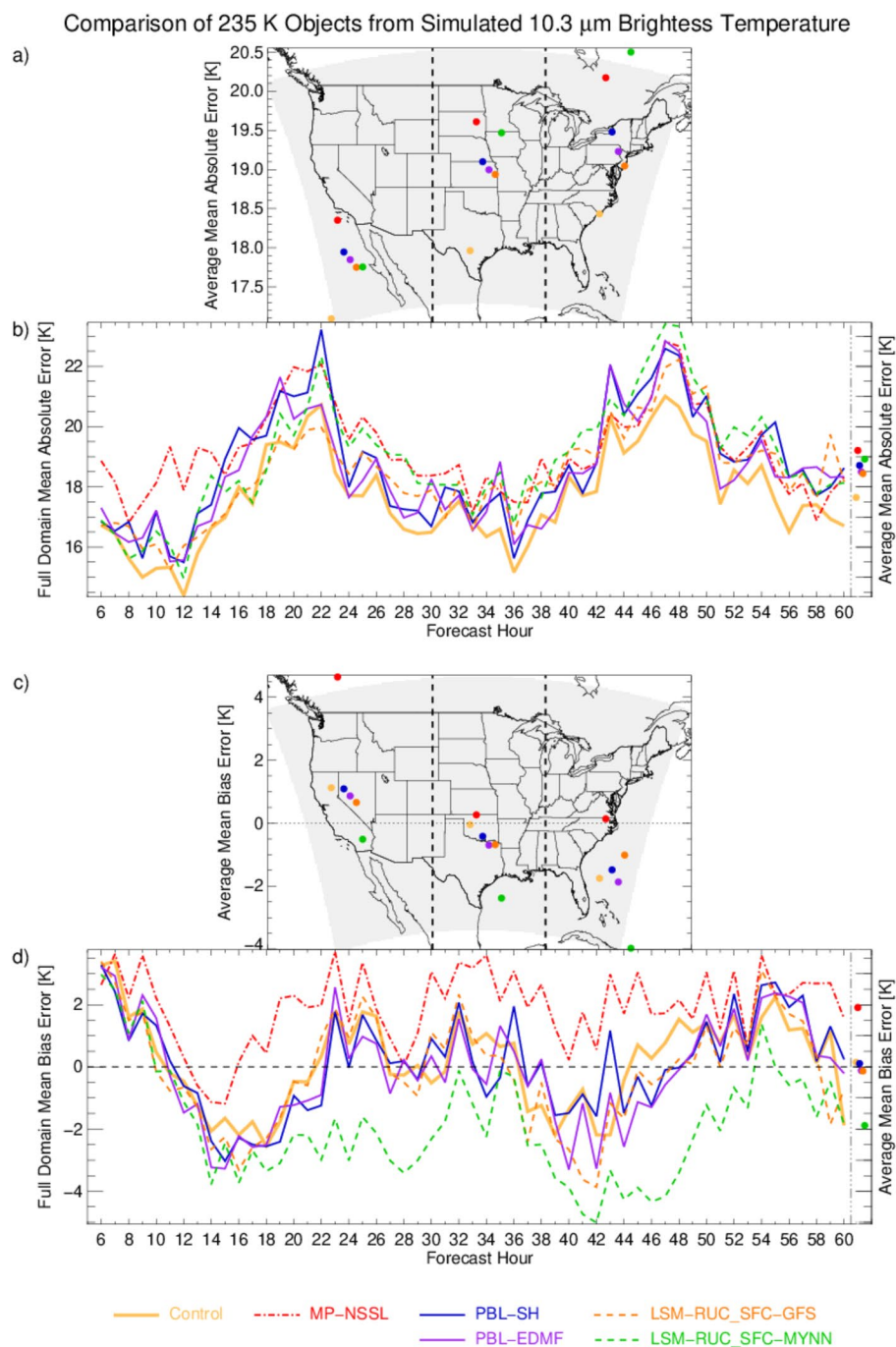
Figure 10. (a) Average mean absolute error (MAE) and (c) mean bias error (MBE) for each sector of the contiguous United States over the full forecast period. Line plot of (b) MAE and (d) MBE for the cloud objects containing ABI 10.3 μm brightness temperatures lower than 235 K as identified by Method for Object-Based Diagnostic Evaluation based on forecast hour. Objects are overlaid using the method from Figure 3. EDMF, hybrid eddy-diffusivity mass-flux; GFS, Global Forecasting System; LSM, land surface models; MYNN, Mellor-Yamanda-Nakanishi-Niino; PBL, planetary boundary layer; SH, Shin-Hong.

analysis of the object attributes (not shown) revealed that there is a local maximum in centroid distance interest scores at the same time as the local maximum in OTS due to smaller displacement between object pairs. This local maximum is not observed in the area ratio interest scores.

Table 3
Average MAE and MBE for 235 K Objects With Displacement Removed

Experiment	MAE (K)	MBE (K)
Control	17.65	0.15
MP-NSSL	19.20	1.91
PBL-SH	18.71	0.11
PBL-EDMF	18.50	−0.12
LSM-RUC_SFC-GFS	18.43	−0.13
LSM-RUC_SFC-MYNN	18.92	−1.89

Abbreviations: EDMF, hybrid eddy-diffusivity mass-flux; GFS, Global Forecasting System; LSM, land surface models; MAE, mean absolute error; MBE, mean bias error; MYNN, Mellor-Yamanda-Nakanishi-Niino; PBL, planetary boundary layer; SH, Shin-Hong.

4.4. Accounting for BT Biases in the Model Configurations

The results presented in previous sections have revealed that BT biases exist in the forecasts and that they differ depending on which microphysics, PBL, and LSM are used. Since accounting for these biases when identifying the cloud objects could result in different performance for each configuration, cloud objects are also defined using a dynamic threshold based on the 6.5th percentile of the BT distribution for each model configuration. The 6.5th percentile is chosen as this is 235 K in the observed BTs. This framework removes the influence of coverage biases on the error statistics and quantifies potential forecast improvements that could be made through a more accurate representation of the spatial extent of upper-level clouds. The 6.5th percentile was identified using the cumulative distribution function of BTs for the observations and each model configuration using all forecasts starting with FH 6, as seen in Figure 11. Overall, all of the model configurations have BTs < 235 K at the 6.5th percentile (Table 4).

A comparison of the OTS for objects identified using a BT threshold equal to either 235 K or the 6.5th percentile for each configuration is shown in Figure 12. It is evident that differences in Figure 12a are neutral overall, with the average percent change from a threshold of 235 K to the 6.5th percentile being approximately zero. Control still has the highest OTS when using the 6.5th percentile of BTs to define the cloud objects (Figure 12b) just like when using a threshold of 235 K (Figure 12c). However, only MP-NSSL has an average OTS that is higher when defining objects using the 6.5th percentile of BTs instead of 235 K. MED for objects defined with the 6.5th percentile of BTs and 235 K has similar results to the OTS, with Control having the lowest MED from forecast object grid points to observations and observation object grid point to forecast (not shown).

For the MAE, Control again has the lowest MAE (Figure 13a) for each sector and over the full domain. Overall, the relative performance of each model configuration based on MAE is unchanged. For the MBE, one noticeable difference exists when evaluating objects defined with the 6.5th percentile or 235 K threshold. While the results are still consistent with MP-NSSL having the highest MBE, model configurations now have a lower MBE as seen in Figure 13b. Overall, the number of pixels used when calculating the MAE and
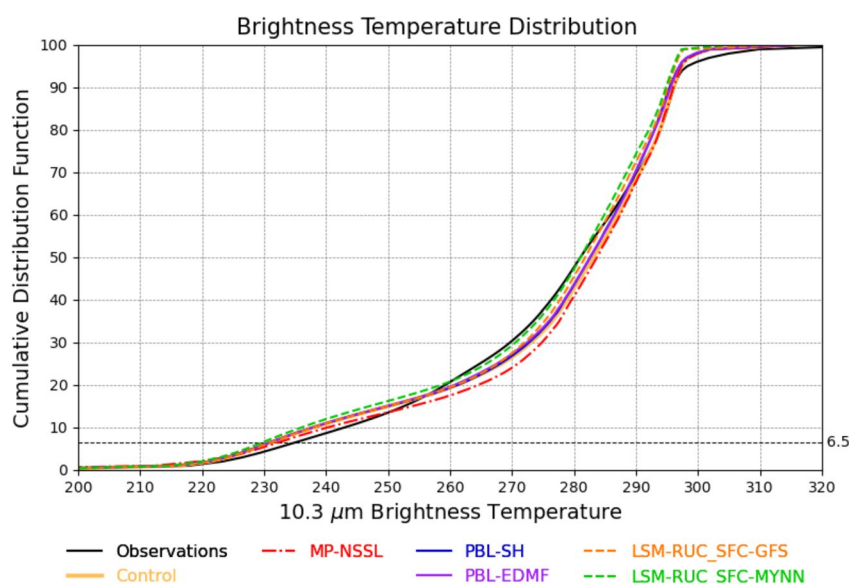

Figure 11. Cumulative distribution of 10.3 μm brightness temperatures (BTs). The 6.5th percentile is highlighted as it represents an observation BT of 235.0 K. EDMF, hybrid eddy-diffusivity mass-flux; GFS, Global Forecasting System; LSM, land surface models; MYNN, Mellor-Yamanda-Nakanishi-Niino; PBL, planetary boundary layer; SH, Shin-Hong.

Table 4
The 6.5th Percentile of Simulated BTs for Each Model Configuration

Configuration	6.5th BT threshold (K)
Control	231.0
MP-NSSL	232.3
PBL-SH	230.9
PBL-EDMF	230.9
LSM-RUC_SFC-GFS	231.1
LSM-RUC_SFC-MYNN	229.7

Note. The 6.5th percentile represents 235.0 K in the observations. Abbreviations: BT, brightness temperature; EDMF, hybrid eddy-diffusivity mass-flux; GFS, Global Forecasting System; LSM, land surface models; MYNN, Mellor-Yamanda-Nakanishi-Niino; PBL, planetary boundary layer; SH, Shin-Hong.

MBE for the 6.5th percentile of BTs is smaller than when using the 235 K threshold, and the objects now encompass more of the analysis box. Therefore, higher BTs surrounding the objects were potentially responsible for the higher MBEs when using an object BT threshold of 235 K.

5. Discussion and Conclusions

In this study, object-based verification is used to assess the accuracy of clouds predicted in the upper troposphere in FV3-LAM simulations performed during the 2019 HWT Spring Forecasting Experiment. This assessment was accomplished by comparing MODE-identified objects in observed and simulated GOES-16 ABI 10.3 μm BTs. Six model configurations employing different LSM and microphysics, PBL, and surface layer schemes, were evaluated. The forecast accuracy was assessed using the (a) OTS that quantifies the similarity of matched cloud objects using MODE interest scores and the ratio of their areas to the total area of all matched and unmatched objects, (b) MED, which calculates the distance between every grid point in a forecast cloud object to the closest observation object grid point, and vice versa, and (c) pixel-based metrics such as the MAE and MBE which assess the BTs of the matched object pairs. The MAE and MBE are only calculated for observation objects with a matching object in all six of the corresponding model forecasts and uses the BTs from grid points in the object and surrounding area after spatial displacement between the objects is removed. Therefore, these methods assess how well the objects' shape and size match, the displacement between objects, and how well the simulated BTs from matched forecast objects and the surrounding area compare to the observed BTs.

Based on the OTS and MAE, over the full forecast period the configuration with the Thompson aerosol-aware microphysics scheme, the MYNN PBL, the GFS surface layer, the Noah LSM, and the NAM initial
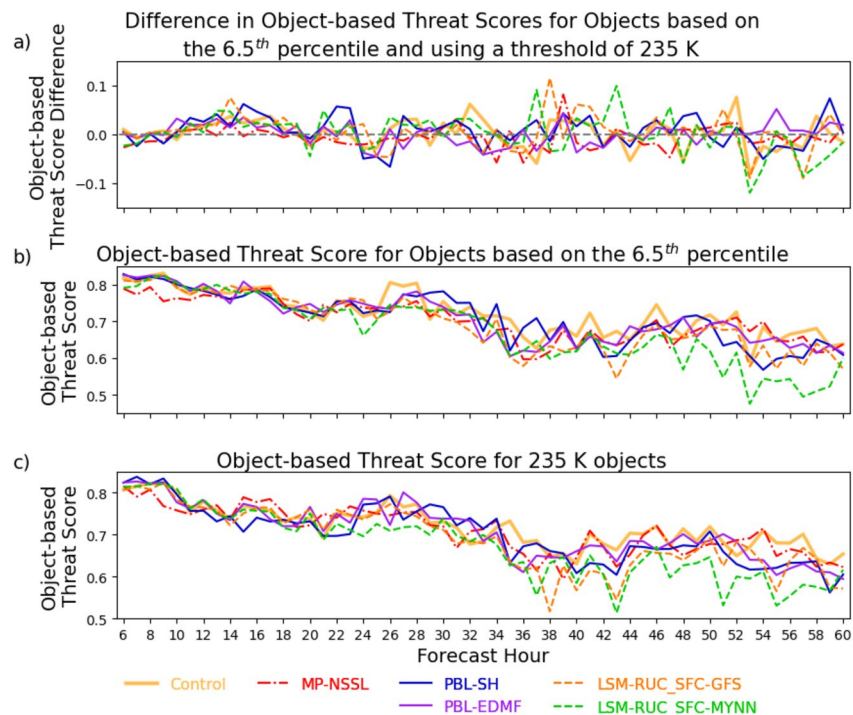

Figure 12. (a) Difference in object-based threat score when defining cloud objects using the 6.5th percentile of brightness temperature distribution (b) and with a fixed threshold of 235 K (c). EDMF, hybrid eddy-diffusivity mass-flux; GFS, Global Forecasting System; LSM, land surface models; MYNN, Mellor-Yamanda-Nakanishi-Niino; PBL, planetary boundary layer; SH, Shin-Hong.
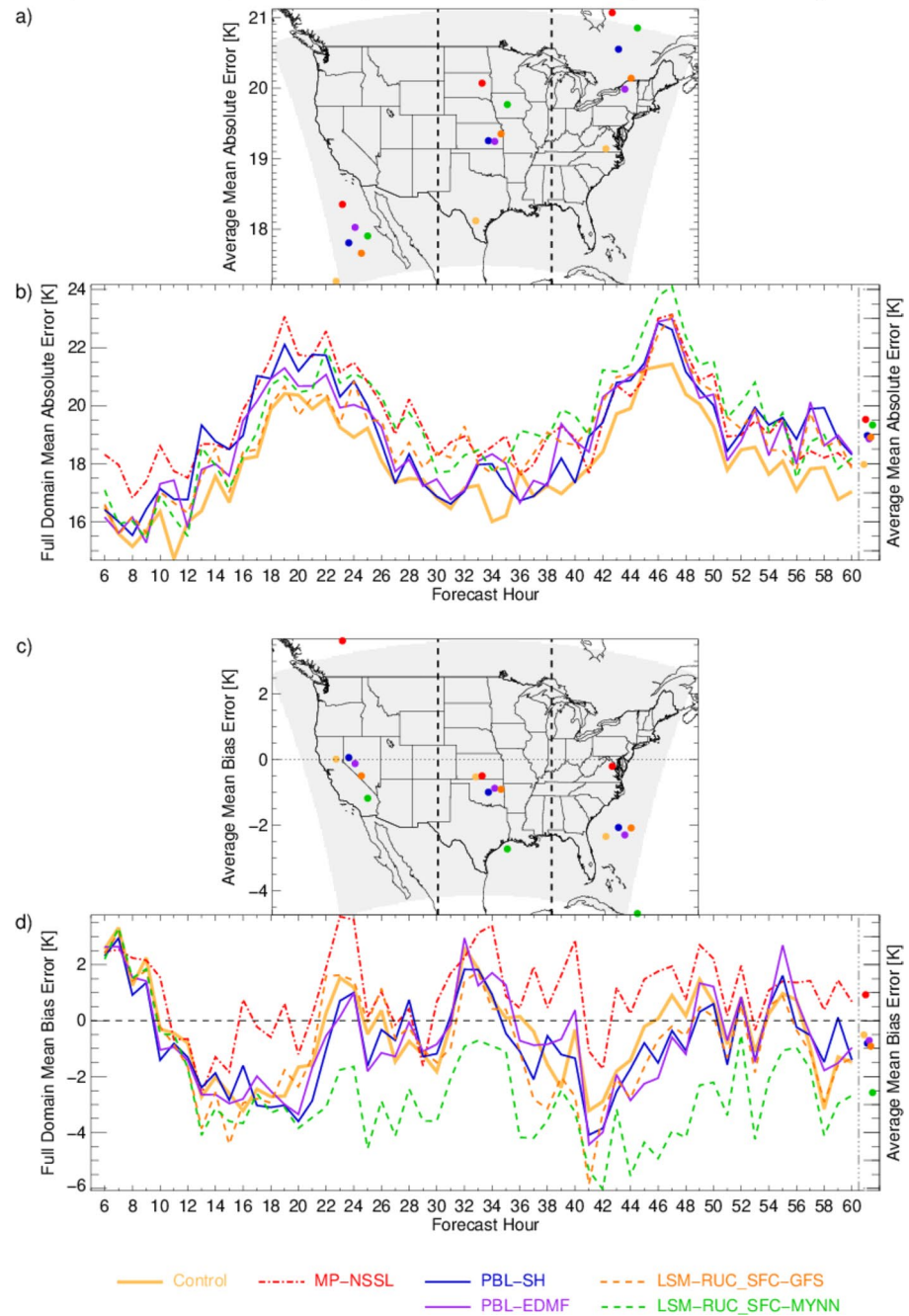
Figure 13. Same as Figure 10 but for objects defined using the 6.5th percentile of brightness temperatures. EDMF, hybrid eddy-diffusivity mass-flux; GFS, Global Forecasting System; LSM, land surface models; MYNN, Mellor-Yamanda-Nakanishi-Niino; PBL, planetary boundary layer; SH, Shin-Hong.

and boundary conditions was the most accurate model configuration. This result is true no matter whether the cloud objects are defined using a BT threshold of 235 K or the 6.5th percentile of the BT distribution for the observations and each configuration. Changing the microphysics scheme from Thompson to NSSL resulted in less accurate cloud objects due to an increase in the median number and area covered by cloud objects, which results in a lower percent of forecast cloud area being matched and a lower OTS. These results are consistent with Arbizu-Barrena et al. (2015) who found that NSSL produced a higher frequency

bias than the Thompson scheme when comparing cloud cover to ceilometer data. Interestingly, for matched object pairs over the full CONUS domain, MP-NSSL had a higher MBE in the BTs than occurred when using the Thompson MP scheme. Therefore, it appears that the inclusion of graupel particle density and the mass and number of hail particles in the NSSL microphysics scheme (which are not in the Thompson scheme) acts to increase the BTs. Since changes in cloud ice have the largest impact on simulated IR BTs (Grasso & Greenwald, 2004), it is possible the NSSL MP scheme has less cloud ice than the Thompson due to the hail category. In addition, forecast objects in matched object pairs when using the MP-NSSL scheme are generally smaller than objects from the Thompson MP scheme. These smaller objects could be due to less cloud water, as Cohen and McCaul (2006) indicated cloud water is collected by graupel and higher cloud water resulted in larger anvils. Fewer large-scale clouds could result in more convective activity, as noted by Tost et al. (2007). However, future work is needed to confirm if using the NSSL MP scheme results in a smaller amount of cloud ice or cloud water than the Thompson MP scheme.

Changing the PBL scheme from the MYNN to the Shin-Hong or EDMF schemes slightly reduces the overall forecast accuracy based on the lower OTS and higher MAE over the full forecast length. However, it is important to note that using the Shin-Hong PBL scheme results in a higher OTS for FHs 6–24 than the MYNN PBL scheme because its objects match more of the observation object during these FHs. One reason for the lower OTS is lower average interest scores, partly due to lower average area ratio interest scores from slightly larger objects. This increase in object area also increases the MED, though this increase compared to using the MYNN PBL scheme is only ∼1 grid point. These larger objects also result in BTs that are lower for the object and surrounding area for both thresholds used to define the cloud objects, and therefore the BTs are less accurate. The higher MAE for the Shin-Hong PBL scheme compared to MYNN is consistent with the findings in Cintineo et al. (2014) that the Thompson MP scheme with Yonsei University PBL scheme, which the Shin-Hong is based on (M. Huang et al., 2019), had a slightly higher MAE than the Thompson MP scheme with the MYNN PBL. It is difficult to say why one PBL scheme is more accurate than another in this analysis. Shin and Dudhia (2016) found that vertical velocities for an EDMF PBL better represented a reference data set than the MYNN PBL, which could be why PBL-EDMF has a lower median number of objects when using a BT threshold of 235 K and better represents the number of observation objects. However, the EDMF PBL overestimates the area covered by objects compared to the MYNN PBL. This could be due to a higher entrainment ratio in the MYNN PBL scheme compared to the EDMF (Shin & Dudhia, 2016), as higher entrainment is found to reduce clouds (Lamraoui et al., 2019).

For the configurations using different LSMs and surface layers, the impact of the different surface features becomes more pronounced with increasing FH. One potential reason for this is the RUC LSM was found to have a higher surface air temperature than the Noah LSM (J. Jin et al., 2010) which could lead to more development of upper-level clouds. LSMs impact the surface air temperature through changes in the partitioning of net surface radiation into sensible and latent heat fluxes (Friedl, 2002). Over the full forecast time period, both LSM-RUC_SFC-GFS and LSM-RUC_SFC-MYNN have lower MBE and higher MAE than the Noah LSM and GFS surface layer, regardless of whether a threshold of 235 K or one based on the 6.5th percentile of the BT distribution is used to identify the cloud objects. They also have a slightly lower OTS due to a smaller percentage of forecast and observation area matched and lower average interest scores, and have a slightly higher MED(O, F). Changing the surface layer from GFS to MYNN further leads to lower BTs and increases the number of objects more rapidly as the forecast progresses.

As this study focuses solely at the 10.3 μm BTs, future work includes extending this analysis to other GOES-16 ABI bands. For example, the 6.2 and 6.9 μm bands are sensitive to water vapor at different levels of the troposphere and can be used to examine other atmospheric features such as jet streams and troughs. Other work will analyze radar data in conjunction with satellite-based object verification to gain a deeper understanding of the forecast accuracy.

Data Availability Statement

References

Arbizu-Barrena, C., Pozo-Vázquez, D., Ruiz-Arias, J. A., & Tovar-Pescador, J. (2015). Macroscopic cloud properties in the WRF NWP model: An assessment using sky camera and ceilometer data. *Journal of Geophysical Research: Atmospheres*, *120*, 10297–10312. https://doi.org/10.1002/2015JD023502

Benjamin, S., Weygandt, S., Brown, J., Hu, M., Alexander, C., Smirnova, T., et al. (2016). A North American hourly assimilation and model forecast cycle: The rapid refresh. *Monthly Weather Review*, *144*(4), 1669–1694. https://doi.org/10.1175/mwrd-15-0242.110.1175/mwr-d-15-0242.1

Bikos, D., Lindsey, D. T., Otkin, J., Sieglaff, J., Grasso, L., Siewert, C., et al. (2012). Synthetic satellite imagery for real-time high-resolution model evaluation. *Weather and Forecasting*, *27*(3), 784–795. https://doi.org/10.1175/WAF-D-11-00130.1

Bullock, R. G., Brown, B. G., & Fowler, T. L. (2016). *Method for object-based diagnostic evaluation* (No. NCAR/TN-532+STR). https://doi.org/10.5065/D61V5CBS

Bytheway, J. L., & Kummerow, C. D. (2015). Toward an object-based assessment of high-resolution forecasts of long-lived convective precipitation in the central U.S. *Journal of Advances in Modeling Earth Systems*, *7*(3), 1248–1264. https://doi.org/10.1002/2015ms000497

Cai, H., & Dumais, R. E., Jr. (2015). Object-based evaluation of a numerical weather prediction model's performance through forecast storm characteristic analysis. *Weather and Forecasting*, *30*(6), 1451–1468. https://doi.org/10.1175/WAF-D-15-0008.1

Chang, F.-L., & Li, Z. (2005). A near-global climatology of single-layer and overlapped clouds and their optical properties retrieved from Terra/MODIS data using a new algorithm. *Journal of Climate*, *18*(22), 4752–4771. https://doi.org/10.1175/JCLI3553.1

Cintineo, J. L., Pavolonis, M. J., Sieglaff, J. M., & Heidinger, A. K. (2013). Evolution of severe and nonsevere convection inferred from GOES-derived cloud properties. *Journal of Applied Meteorology and Climatology*, *52*(9), 2009–2023. https://doi.org/10.1175/JAMC-D-12-0330.1

Cintineo, J. L., Smith, T. M., Lakshmanan, V., Brooks, H. E., & Ortega, K. L. (2012). An objective hail climatology of the contiguous United States. *Weather and Forecasting*, *27*(5), 1235–1248. https://doi.org/10.1175/WAF-D-11-00151.1

Cintineo, R., Otkin, J. A., Kong, F., & Xue, M. (2014). Evaluating the accuracy of planetary boundary layer and cloud microphysical parameterization schemes in a convection-permitting ensemble using synthetic GOES-13 satellite observations. *Monthly Weather Review*, *142*(1), 107–182. https://doi.org/10.1175/MWR-D-13-00143.1

Clark, A. J., Jirak, I. L., Gallo, B. T., Roberts, B., Knopfmeier, K. H., Clark, R. A., et al. (2020). A real-time, simulated forecasting experiment for advancing the prediction of hazardous convective weather. *Bulletin of the American Meteorological Society*, *101*(11), E2022–E2024. https://doi.org/10.1175/BAMS-D-19-0298.1

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

Cohen, C., & McCaul, E. W., Jr. (2006). The sensitivity of simulated convective storms to variations in prescribed single-moment microphysics parameters that describe particle distributions, sizes, and numbers. *Monthly Weather Review*, *134*(9), 2547–2565. https://doi.org/10.1175/MWR3195.1

Davis, C. A., Brown, B. G., & Bullock, R. G. (2006a). 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. A., Brown, B. G., & Bullock, R. G. (2006b). Object-based verification of precipitation forecasts. Part II: Application to convective rain systems. *Monthly Weather Review*, *134*(7), 1785–1795. https://doi.org/10.1175/MWR3146.1

Davis, C. A., Brown, B. G., Bullock, R. G., & 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

Ding, S., Yang, P., Weng, F., Liu, Q., Han, Y., Van Delst, P., et al. (2011). Validation of the community radiative transfer model. *Journal of Quantitative Spectroscopy and Radiative Transfer*, *112*(6), 1050–1064. https://doi.org/10.1016/j.jqsrt.2010.11.009

Federal Aviation Administration. (2018). *FAA aerospace forecast: Fiscal years 2019-2039*. Retrieved from https://www.faa.gov/data_research/aviation/aerospace_forecasts/media/FY2019-39_FAA_Aerospace_Forecast.pdf

Feltz, W. F., Bedka, K. M., Otkin, J. A., Greenwald, T., & Ackerman, S. A. (2009). Understanding satellite-observed mountain wave signatures using high-resolution numerical model data. *Weather and Forecasting*, *24*(1), 76–86. https://doi.org/10.1175/2008WAF2222127.1

Friedl, M. A. (2002). Forward and inverse modeling of land surface energy balance using surface temperature measurements. *Remote Sensing of Environment*, *79*(2–3), 344–354. https://doi.org/10.1016/S0034-4257(01)00284-X

Gallo, B. T., Clark, A. J., Jirak, I., Kain, J. S., Weiss, S. J., Coniglio, M., et al. (2017). Breaking new ground in severe weather prediction: The 2015 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. *Weather and Forecasting*, *32*(4), 1541–1568. https://doi.org/10.1175/WAF-D-16-0178.1

Gilleland, E. (2017). A new characterization in the spatial verification framework for false alarms, misses, and overall patterns. *Weather and Forecasting*, *32*(1), 187–198. https://doi.org/10.1175/WAF-D-16-0134.1

Gilleland, E., Ahijevych, D., Brown, B. G., Casati, B., & Ebert, E. E. (2009). Intercomparison of spatial forecast verification methods. *Weather and Forecasting*, *24*(5), 1416–1430. https://doi.org/10.1175/2009WAF2222269.1

Gilleland, E., Ahijevych, D., Brown, B. G., & Ebert, E. E. (2010). Verifying forecasts spatially. *Bulletin of the American Meteorological Society*, *91*(10), 1365–1376. https://doi.org/10.1175/2010BAMS2819.1

Grasso, L. D., & Greenwald, T. (2004). Analysis of 10.7-μm brightness temperatures of a simulated thunderstorm with two-moment microphysics. *Monthly Weather Review*, *132*(3), 815–825. https://doi.org/10.1175/1520-0493(2004)132<0815:aombto>2.0.co;2

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

Grasso, L. D., Sengupta, M., Dostalek, J. F., Brummer, R., & DeMaria, M. (2008). Synthetic satellite imagery for current and future environmental satellites. *International Journal of Remote Sensing*, *29*(15), 4373–4384. https://doi.org/10.1080/01431160801891820

Griffin, S. M., Otkin, J. A., Rozoff, C. M., Sieglaff, J. M., Cronce, L. M., & Alexander, C. R. (2017a). Methods for comparing simulated and observed satellite infrared brightness temperatures and what do they tell us? *Weather and Forecasting*, *32*(1), 5–25. https://doi.org/10.1175/WAF-D-16-0098.1

Griffin, S. M., Otkin, J. A., Rozoff, C. M., Sieglaff, J. M., Cronce, L. M., Alexander, C. R., et al. (2017b). 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

Griffin, S. M., Otkin, J. A., Thompson, G., Frediani, M., Berner, J., & Kong, F. (2020). Assessing the impact of stochastic perturbations in cloud microphysics using GOES-16 infrared brightness temperatures. *Monthly Weather Review*, *148*(8), 3111–3137. https://doi.org/10.1175/MWR-D-20-0078.1

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., et al. (2005). *JCSDA community radiative transfer model (CRTM) version 1* (NOAA Technical Report 122). National Environmental Satellite, Data, and Information Service, Silver Spring.

Haupt, S. E., Kosovic, B., Jensen, T., Lazo, J. K., Lee, J. A., Jimenez, P. A., et al. (2018). Building the Sun4Cast system: Improvements in solar power forecasting. *Bulletin of the American Meteorological Society*, *99*(1), 121–136. https://doi.org/10.1175/BAMS-D-16-0221.1

Huang, M., Gao, Z., Miao, S., & Chen, F. (2019). Sensitivity of urban boundary layer simulation to urban canopy models and PBL schemes in Beijing. *Meteorology and Atmospheric Physics*, *131*, 1235–1248. https://doi.org/10.1007/s00703-018-0634-1

Huang, Y., & Peng, X. (2017). Improvement of the Mellor–Yamada–Nakanishi–Niino planetary boundary-layer scheme based on observational data in China. *Boundary-Layer Meteorology*, *162*, 171–188. https://doi.org/10.1007/s10546-016-0187-0

Jankov, I., Grasso, L., Sengupta, M., Neiman, P. J., Zupanski, D., Lindsey, D., et al. (2011). An evaluation of five ARW-WRF microphysics schemes using synthetic GOES imagery for an atmospheric river event affecting the California Coast. *Journal of Hydrometeorology*, *12*(4), 618–633. https://doi.org/10.1175/2010JHM1282.1

Jensen, T., Brown, B., Bullock, R., Fowler, T., Halley Gotway, J., & Newman, K. (2020). *Model evaluation tools version 9.0.1 user's guide*. Retrieved from https://dtcenter.org/sites/default/files/community-code/met/docs/user-guide/MET_Users_Guide_v9.0.1.pdf

Jin, H., Peng, M. S., Jin, Y., & Doyle, J. D. (2014). An evaluation of the impact of horizontal resolution on tropical cyclone predictions using COAMPS-TC. *Weather and Forecasting*, *29*(2), 252–270. https://doi.org/10.1175/WAF-D-13-00054.1

Jin, J., Miller, N. L., & Schegel, N. (2010). Sensitivity study of four land surface schemes in the WRF model. *Advances in Meteorology*, 167436. https://doi.org/10.1155/2010/167436

Johnson, A., & Wang, X. (2013). Object-based evaluation of a storm-scale ensemble during the 2009 NOAA Hazardous Weather Testbed Spring Experiment. *Monthly Weather Review*, *141*(3), 1079–1098. https://doi.org/10.1175/MWR-D-12-00140.1

Johnson, R., & Mapes, B. E. (2001). Mesoscale processes and severe convective weather. In *Severe convective storms* (Meteorological Monographs No. 50, pp. 71–122). American Meteorological Society. https://doi.org/10.1007/978-1-935704-06-5_3

Jolliffe, I. T., & Stephenson, D. B. (2012). *Forecast verification: A practioner's Guide in atmospheric science* (2nd ed., p. 274). Wiley-Blackwell.

Jones, T. A., Skinner, P., Knopfmeier, K., Mansell, E., Minnis, P., Palikonda, R., & Smith, W. (2018). Comparison of cloud microphysics schemes in a Warn-On-Forecast system using synthetic satellite objects. *Weather and Forecasting*, *33*(6), 1681–1708. https://doi.org/10.1175/WAF-D-18-0112.1

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

Kaplan, M. L., Huffman, A. W., Lux, K. M., Charney, J. J., Riordan, A. J., & Lin, Y.-L. (2005). Characterizing the severe turbulence environments associated with commercial aviation accidents. Part 1: A 44-case study synoptic observational analyses. *Meteorology and Atmospheric Physics*, *88*, 129–152. https://doi.org/10.1007/s00703-004-0080-0

Lamraoui, F., Booth, J. F., Naud, C. M., Jensen, M. P., & Johnson, K. L. (2019). The interaction between boundary layer and convection schemes in a WRF simulation of post cold frontal clouds over the ARM East North Atlantic site. *Journal of Geophysical Research: Atmospheres*, *124*, 4699–4721. https://doi.org/10.1029/2018jd029370

Lee, J. A., Haupt, S. E., Jiménez, P. A., Rogers, M. A., Miller, S. D., & McCandless, T. C. (2017). Solar irradiance nowcasting case studies near Sacramento. *Journal of Applied Meteorology and Climatology*, *56*(1), 85–108. https://doi.org/10.1175/JAMC-D-16-0183.1

Lin, S. (2004). A "vertically lagrangian" finite-volume dynamical core for global models. *Monthly Weather Review*, *132*(10), 2293–2307. https://doi.org/10.1175/1520-0493(2004)132<2293:avlfdc>2.0.co;2

Mansell, E. R., Ziegler, C. L., & Bruning, E. C. (2010). Simulated electrification of a small thunderstorm with two-moment bulk microphysics. *Journal of the Atmospheric Sciences*, *67*(1), 171–194. https://doi.org/10.1175/2009JAS2965.1

Mecikalski, J. R., & Bedka, K. M. (2006). Forecasting convective initiation by monitoring the evolution of moving cumulus in daytime GOES imagery. *Monthly Weather Review*, *134*(1), 49–78. https://doi.org/10.1175/MWR3062.1

Mecikalski, J. R., Feltz, W. F., Murray, J. J., Johnson, D. B., Bedka, K. M., Bedka, S. T., et al. (2007). Aviation applications for satellite-based observations of cloud properties, convective initiation, in-flight icing, turbulence, and volcanic ash. *Bulletin of the American Meteorological Society*, *88*(10), 1589–1607. https://doi.org/10.1175/BAMS-88-10-1589

Mitchell, K. (2005). The community Noah land-surface model. In *User's guide public release version 2.7.1* (p. 26). Environmental Modeling Center, NOAA/NCEP.

Mittermaier, M. P., & Bullock, R. (2013). Using MODE to explore the spatial and temporal characteristics of cloud cover forecasts from high-resolution NWP models. *Meteorological Applications*, *20*(2), 187–196. https://doi.org/10.1002/met.1393

Mittermaier, M. P., North, R., Semple, A., & Bullock, R. (2016). Feature-based diagnostic evaluation of global NWP forecasts. *Monthly Weather Review*, *144*(10), 3871–3893. https://doi.org/10.1175/MWR-D-15-0167.1

Morcrette, J. J. (1991). Evaluation of model-generated cloudiness: Satellite-observed and model-generated diurnal variability of brightness temperature. *Monthly Weather Review*, *119*, 1205–1224. https://doi.org/10.1175/1520-0493(1991)119<1205:eomgcs>2.0.co;2

Murray, J. J. (2002). Aviation weather applications of Earth Science Enterprise data. *Earth Observation Magazine*, *11*(8), 26–30.

Nakanishi, M., & Niino, H. (2004). An improved Mellor–Yamada level-3 model with condensation physics: Its design and verification. *Boundary-Layer Meteorology*, *112*, 1–31. https://doi.org/10.1023/B:BOUN.0000020164.04146.98

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

Niu, G.-Y. (2011). *The community Noah land-surface model (LSM) with multi-physics options*. User's Guide Public.

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

Otkin, J. A., Greenwald, T. J., Sieglaff, J., & Huang, H.-L. (2009). Validation of a large-scale simulated brightness temperature dataset using SEVIRI satellite observations. *Journal of Applied Meteorology and Climatology*, *48*(8), 1613–1626. https://doi.org/10.1175/2009JAMC2142.1

Otkin, J. A., Lewis, W. E., Lenzen, A., McNoldy, B., & Majumdar, S. (2017). Assessing the accuracy of the cloud and water vapor fields in the Hurricane WRF (HWRF) model using satellite infrared brightness temperatures. *Monthly Weather Review*, *145*(5), 2027–2046. https://doi.org/10.1175/MWR-D-16-0354.1

Otkin, J. A., Posselt, D. J., Olson, E. R., Huang, H.-L., Davies, J. E., Li, J., & Velden, C. S. (2007). Mesoscale numerical weather prediction models used in support of infrared hyperspectral measurement simulation and product algorithm development. *Journal of Atmospheric and Oceanic Technology*, *24*(4), 585–601. https://doi.org/10.1175/JTECH1994.1

Purdom, J. F. W. (1993). Satellite observations of tornadic thunderstorms. In *The Tornado: Its structure, dynamics, prediction, and hazards* (Geophysical Monograph Series, Vol. 79, pp. 265–274). American Geophysical Union. https://doi.org/10.1029/gm079p0265

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

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

Senf, F., Klocke, D., & Brueck, M. (2018). Size-resolved evaluation of simulated deep tropical convection. *Monthly Weather Review*, *146*(7), 2161–2182. https://doi.org/10.1175/MWR-D-17-0378.1

Shin, H. H., & Dudhia, J. (2016). Evaluation of PBL parameterizations in WRF at subkilometer grid spacings: Turbulence statistics in the dry convective boundary layer. *Monthly Weather Review*, *144*(3), 1161–1177. https://doi.org/10.1175/MWR-D-15-0208.1

Shin, H. H., & Hong, S.-Y. (2013). Analysis of resolved and parameterized vertical transports in convective boundary layers at gray-zone resolutions. *Journal of the Atmospheric Sciences*, *70*(10), 3248–3261. https://doi.org/10.1175/JAS-D-12-0290.1

Sieglaff, J. M., Cronce, L. M., Feltz, W. F., Bedka, K. M., Pavolonis, M. J., & Heidinger, A. K. (2011). Nowcasting convective storm initiation using satellite-based box-averaged cloud-top cooling and cloud-type trends. *Journal of Applied Meteorology and Climatology*, *50*(1), 110–126. https://doi.org/10.1175/2010JAMC2496.1

Skinner, P. S., Wheatley, D. M., Knopfmeier, K. H., Reinhart, A. E., Choate, J. J., Jones, T. A., et al. (2018). Object-based verification of a prototype Warn-on-Forecast system. *Weather and Forecasting*, *33*(5), 1225–1250. https://doi.org/10.1175/WAF-D-18-0020.1

Skinner, P. S., Wicker, L. J., Wheatley, D. M., & Knopfmeier, K. H. (2016). Application of two spatial verification methods to ensemble forecasts of low-level rotation. *Weather and Forecasting*, *31*(3), 713–735. https://doi.org/10.1175/WAF-D-15-0129.1

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

Sušelj, K., & Sood, A. (2010). Improving the Mellor–Yamada–Janjić Parameterization for wind conditions in the marine planetary boundary layer. *Boundary-Layer Meteorology*, *136*, 301–324. https://doi.org/10.1007/s10546-010-9502-3

Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research*, *106*(D7), 7183–7192. https://doi.org/10.1029/2000JD900719

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

Thompson, G., Rasmussen, R. M., & Manning, K. (2004). Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part I: Description and sensitivity analysis. *Monthly Weather Review*, *132*(2), 519–542. https://doi.org/10.1175/1520-0493(2004)132<0519:efowpu>2.0.co;2

Thompson, G., Tewari, M., Ikeda, K., Tessendorf, S., Weeks, C., Otkin, J. A., & Kong, F. (2016). Explicitly-coupled cloud physics and radiation parameterizations and subsequent evaluation in WRF high-resolution convective forecasts. *Atmospheric Research*, *168*(1), 92–104. https://doi.org/10.1016/j.atmosres.2015.09.005

Tost, H., Jöckel, P., & Lelleveld, J. (2007). Lightning and convection parameterisations—Uncertainties in global modelling. *Atmospheric Chemistry and Physics*, *7*, 4553–4568. https://doi.org/10.5194/acp-7-4553-2007

Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences* (2nd ed., p. 676). Academic Press.

Wolff, J. K., Harrold, M., Fowler, T., Gotway, J. H., Nance, L., & Brown, B. G. (2014). Beyond the basics: Evaluating model-based precipitation forecasts using traditional, spatial, and object-based methods. *Weather and Forecasting*, *29*(6), 1451–1472. https://doi.org/10.1175/WAF-D-13-00135.1

Wong, M., & Skamarock, W. (2017). *Precipitation spin-up processes in a global model with a cloud-permitting-scale mesh refinement* (Vol. 10.4). Preprints, 17th Conference on Mesoscale Processes. American Meteorological Society.

Wu, X., & Schmit, T. (2019). *GOES-16 ABI level1b and cloud and moisture imagery (CMI) release full validation data quality*. Retrieved from http://www.ncdc.noaa.gov/sites/default/files/attachments/GOES-16_ABI-L1b-CMI_Full-Validation_ProductPerformanceGuide_v2.pdf