Cool-Season Evaluation of FV3-LAM-Based CONUS-Scale Forecasts with Physics Configurations of Experimental RRFS Ensembles

TIMOTHY A. SUPINIE,[a] JUN PARK,[a] NATHAN SNOOK,[a] XIAO-MING HU,[a] KEITH A. BREWSTER,[a] MING XUE,[a,b] AND JACOB R. CARLEY[c]

[a] *Center for Analysis and Prediction of Storms, Norman, Oklahoma*
[b] *School of Meteorology, University of Oklahoma, Norman, Oklahoma*
[c] *NOAA/NWS/NCEP/Environmental Modeling Center, College Park, Maryland*

(Manuscript received 16 December 2021, in final form 31 May 2022)

ABSTRACT: To help inform physics configuration decisions and help design and optimize a multi-physics Rapid Refresh Forecasting System (RRFS) ensemble to be used operationally by the National Weather Service, five FV3-LAM-based convection allowing forecasts were run on 35 cases between October 2020 and March 2021. These forecasts used ~3-km grid spacing on a CONUS domain with physics configurations including Thompson, NSSL, and Ferrier–Aligo microphysics schemes, Noah, RUC, and NoahMP land surface models, and MYNN-EDMF, K-EDMF, and TKE-EDMF PBL schemes. All forecasts were initialized from the 0000 UTC GFS analysis and run for 84 h. Also, a subset of 8 cases were run with 15 combinations of physics options, also including the Morrison–Gettelman microphysics and Shin–Hong PBL schemes, to help attribute behaviors to individual schemes and isolate the main contributors of forecast errors. Evaluations of both sets of forecasts find that the CONUS-wide 24-h precipitation > 1 mm is positively biased across all five forecasts. NSSL microphysics displays a low bias in QPF along the Gulf Coast. Analyses show that it produces smaller raindrops prone to evaporation. Additionally, TKE-EDMF PBL in combination with Thompson microphysics displays a positive bias in precipitation over the Great Lakes and in the ocean near Florida due to higher latent heat fluxes calculated over water. Furthermore, the K-EDMF PBL scheme produces temperature errors that result in a negative bias in snowfall over the southern Mountain West. Finally, recommendations for which physics schemes to use in future suites and the RRFS ensemble are discussed.

KEYWORDS: Model errors; Model evaluation/performance; Numerical weather prediction/forecasting; Regional models

1. Introduction

The U.S. National Weather Service (NWS) is in the process of building its entire operational numerical weather prediction (NWP) suite around the nonhydrostatic finite volume cubed-sphere (FV3) dynamical core (Harris et al. 2020; Putman and Lin 2007; Lin 2004), which serves as the foundation of the nation's Unified Forecast System (UFS). According to the plan, the current operational convection-allowing model (CAM) forecasting systems including the North American Mesoscale (NAM) 3-km nest, High-Resolution Rapid Refresh (HRRR), High Resolution Window (HiResW), and High-Resolution Ensemble Forecast (HREF) system will be replaced with a rapidly updated CAM ensemble forecast system called the Rapid Refresh Forecast System (RRFS). The RRFS will use the limited area model version of the FV3 dynamic core (FV3-LAM; Black et al. 2021) and will likely employ more than one optimized suite of physics parameterizations and potentially include stochastic physics perturbations in its forecast ensemble.[1] A multi-physics architecture is chosen for the future RRFS ensemble because a multi-physics ensemble has been found to have superior performance to single-physics ensembles (Berner et al. 2011, 2015), and stochastic physics perturbations have been found to improve the representation of forecast uncertainty (Berner et al. 2015; Jankov et al. 2017). As one of the first steps toward the goal of transitioning to FV3-LAM, candidate physics parameterization schemes implemented within the FV3-based UFS system through the Common Community Physics Package (CCPP; Firl et al. 2021) need to be systematically evaluated for specific applications of the UFS, in our case RRFS forecast ensemble.

NOAA testbed experiments such as the Hazardous Weather Testbed (HWT) Spring Forecasting Experiment (SFE; Clark et al. 2020; Roberts et al. 2020), the Flash Flood and Intense Rainfall (FFaIR) Experiment (Barthold et al. 2015) and the Winter Weather Experiment (WWE; Harnos et al. 2021) of the Hydrometeorology Testbed (HMT) provide ideal environments for running experimental CAM forecasting systems and evaluating the performance of the forecasts, both subjectively by experiment participants in real time and objectively post hoc by forecast users and modelers. The HWT SFE and HMT FFaIR have received more attention so far for spring and summer severe weather and heavy precipitation forecasting, respectively, while winter weather has received less attention, in terms of forecasting performance by CAMs, especially those with new capabilities. For this reason, we choose to focus on evaluation of cool season CAM forecasts run during the 11th HMT WWE.

[1] While the proposed forecast ensemble is a multi-physics ensemble, the proposed data assimilation ensemble is a single-physics ensemble, which avoids the non-Gaussianity inherent in a multi-physics ensemble.

Corresponding author: Timothy A. Supinie, tsupinie@ou.edu

DOI: 10.1175/MWR-D-21-0331.1

Many methods exist for CAM forecast evaluation. One method is to pick a physical process and see how well the model represents that process using observations (e.g., Tessendorf et al. 2021; Ikeda et al. 2013). Another is to define an object in in the forecast grid and compare to objects in an observed or analysis grid (so-called object-based methods). This is the approach taken by, among others, Griffin et al. (2017b), Bytheway et al. (2017), Skinner et al. (2018), Flora et al. (2019), and Duda and Turner (2021). Gridded forecasts can also be evaluated against an analysis grid (Clark et al. 2010; Sobash and Kain 2017; Snook et al. 2019; Sobash et al. 2016). Gallo et al. (2021) compared grid-based methods and object-based methods and found that each illuminated a different aspect of model performance.

Gridpoint-based verification methods can use a variety of statistics and forecast fields for comparison. For accumulated precipitation, simulated radar reflectivity, or updraft helicity fields, statistics based on a 2×2 contingency table are commonly used, such as the critical success index [CSI; e.g., Loken et al. (2017)] and equitable threat score [ETS; e.g., Snook et al. (2019) and Zhang et al. (2019)]. However, gridpoint-based methods are prone to double penalty: for example, penalizing a model which exhibits phase error in a convective line for both having precipitation in the wrong place and not having it in the correct place. To alleviate this, neighborhood methods were developed (e.g., Schwartz et al. 2009). Among these are methods for creating neighborhood-based contingency tables by Clark et al. (2010) and McMillen and Steenburgh (2015). One popular metric developed specifically as a neighborhood metric is the fractions skill score (Roberts and Lean 2008), used by Cintineo et al. (2014) and Griffin et al. (2017a).

Because the application of the FV3 dynamical core at the convective scale is relatively new, many CAM evaluation studies have been based on the Weather Research and Forecasting (WRF) Model as the dynamical core, rather than the relatively new FV3-LAM. To the best of the authors' knowledge, no published literature has examined the performance FV3 on the convective scale during the cool season. However, a few recent studies have examined the performance of FV3 for CAM forecasts, primarily for the warm season (Potvin et al. 2019; Zhang et al. 2019; Griffin et al. 2021). Zhang et al. (2019) evaluated hourly precipitation forecasts from FV3-based CAM forecasts with different physics configurations run during the 2018 HWT SFE. They found that the choice of microphysics had a much larger effect on forecast performance than the choice of planetary boundary layer (PBL) scheme. The Thompson microphysics scheme (Thompson et al. 2004, 2008) slightly outperformed the NSSL microphysics scheme (Mansell et al. 2010; Mansell and Ziegler 2013) in hourly precipitation forecasts, particularly in day-2 and day-3 forecasts, and FV3 overall performed comparably to WRF. Griffin et al. (2021) evaluated simulated brightness temperatures in an object-based framework for FV3-based CAM forecasts with different physics configurations run in spring 2019. They found that Thompson microphysics with the MYNN PBL scheme (Nakanishi and Niino 2009; Olson et al. 2019) performed the best when evaluated on cloud objects, while NSSL microphysics tended to over predict cloud object number and extent. Additionally, changing the PBL scheme from MYNN to Shin–Hong (Shin and Hong 2015) or K-EDMF (Han et al. 2016) generally resulted in slightly lower object accuracy.

The purpose of this study is to evaluate various FV3-LAM physics configurations to help inform the design of the operational RRFS ensemble. While ensembles considered as a whole provide much useful information on convective scales (Roberts et al. 2020), the focus of this study is on the individual physics configurations and their performance. The performance of the ensemble as a whole will be considered in a future study. The remainder of this study is laid out as follows: the physics configurations tested and the methods for testing them are introduced in section 2. Evaluation is presented for surface fields in section 3, precipitation forecasts in section 4, and snowfall forecasts in section 5. Finally, conclusions are discussed in section 6.

2. Forecast configuration and evaluation methods

a. *Forecast configuration*

During the 11th HMT WWE (hereafter referred to as the "HMT forecasts"), CAPS ran five FV3-LAM forecasts per case day. The forecasts were initialized at 0000 UTC on 35 days between October 2020 and March 2021 (see Fig. 1b), which covers the period of the WWE plus additional significant winter weather events in late October 2020 and mid-March 2021. The model domain (Fig. 1a) covers the contiguous United States (CONUS) at approximately 3-km grid spacing with 64 vertical levels. Initial and lateral boundary conditions at 0.25° grid spacing and 3-h time intervals were taken from the corresponding operational run of the NCEP Global Forecast System (GFS) v15 for all forecasts. Different forecasts used different physics configurations (see the italicized rows in Table 1). The microphysics schemes included Thompson, NSSL, and Ferrier–Aligo (Aligo et al. 2018). As part of this work, the NSSL fully two-moment microphysics scheme was added to the CCPP by this research team, thus making it available in FV3-LAM. The PBL schemes used were scale-aware MYNN-EDMF (hereafter referred to as "MYNN"), K-EDMF, and TKE-EDMF (Han and Bretherton 2019). The MYNN PBL scheme was used with its corresponding surface layer scheme (Nakanishi and Niino 2009; Olson et al. 2021), while the K-EDMF and TKE-EDMF were used with the GFS operational surface layer scheme (Long 1986, 1990). The land surface models (LSMs) used were the Noah (Ek et al. 2003), RUC (Smirnova et al. 2016), and NoahMP (Niu et al. 2011) models. All forecasts used the Rapid Radiative Transfer Model for General Circulation Models (RRTMG; Mlawer et al. 1997) as their radiation scheme and the GFS near-surface sea temperature (NSST) scheme, a modified version of Fairall et al. (1996), for latent and sensible heat fluxes over water. In Table 1, the "M," "B," and "L" in the forecast names refer to microphysics, PBL, and LSM, respectively, while the numbers 0, 1, 2, 3 refer to the specific scheme. The schemes used in control forecast M0B0L0 all have "0" designation. The specific physics configurations of the HMT forecasts were

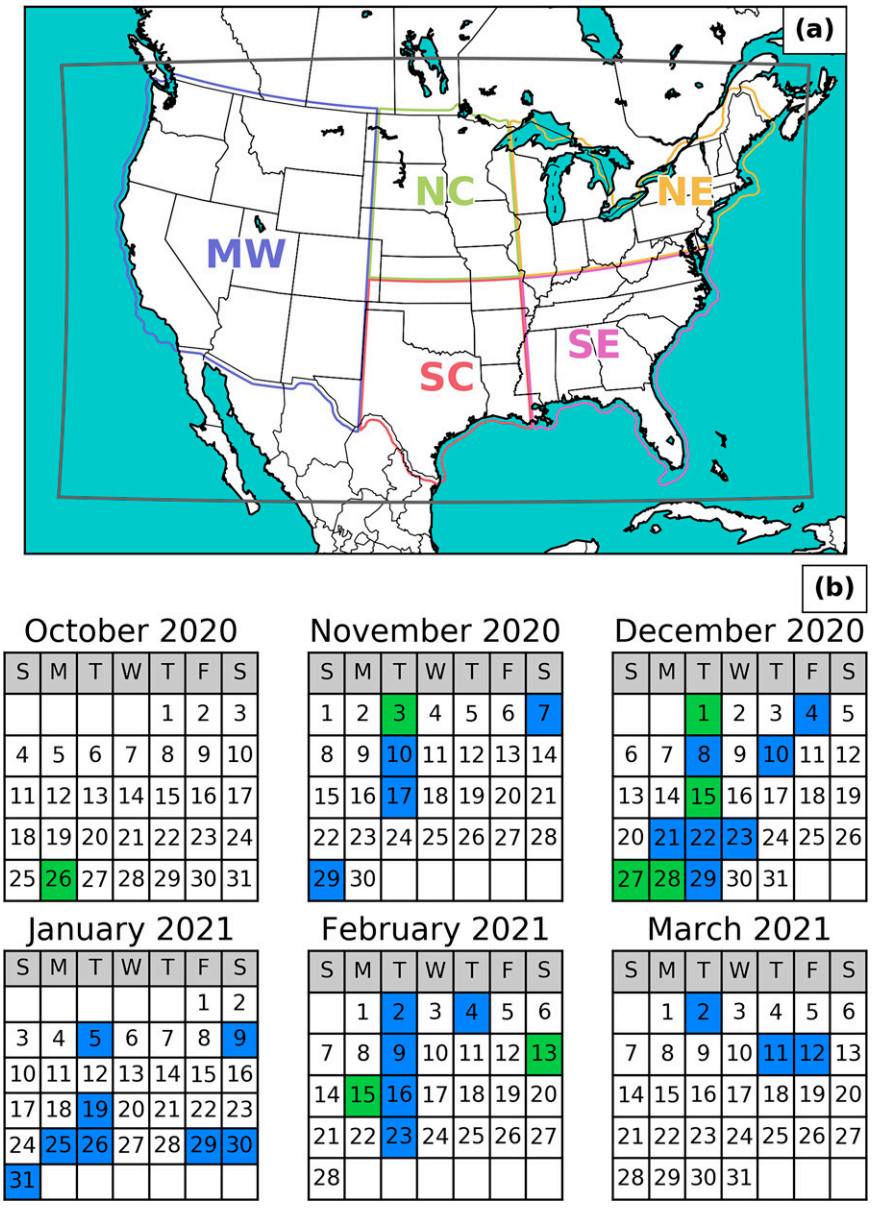

FIG. 1. (a) The native FV3-LAM domain boundary used for all forecast runs is given in gray. The verification subregions are given by the colored outlines: MW, NC, NE, SC, and SE denoting the mountain west, north-central, northeast, south-central, and southeast regions, respectively. The union of all five subregions is the CONUS verification region. (b) Dates for the forecasts. All 15 physics configurations were run on dates highlighted in green. Only the HMT set of physics configurations were run on dates highlighted in blue. All forecasts are initialized at 0000 UTC on their respective dates.

chosen to reflect the current and proposed operational regional model configurations of NWS. Specifically, the control physics configuration M0B0L0 has options similar to those of the proposed control member of operational RRFS, except for the LSM; NoahMP is proposed, but our preliminary evaluations indicate that NoahMP does not perform as well as Noah. Therefore, the latter was chosen for our control forecast. M0B0L2 used the same options as the current operational High-Resolution Rapid Refresh (HRRR) while M1B0L0 used the options of experimental Warn-on-Forecast System (WoFS). M0B2L1 used options planned for the next operational version of NCEP GFS while the options of M3B3L0 resemble those of North American Mesoscale (NAM) model and planned Hurricane Analysis and Forecasting System (HAFS).

To isolate the effects of the individual physics schemes on the forecasts, we ran a set of 15 forecasts on a subset of eight

TABLE 1. Forecast physics configurations. Italicized rows indicate that configuration was run during the 2020–21 HMT WWE. All forecasts use the RRTMG shortwave and longwave radiation scheme, the Cooperative Institute for Research in Environmental Sciences (CIRES) unified gravity wave drag scheme, and the GFS near-surface sea temperature algorithm. In the prototype column, "WoFS" stands for the Warn-on-Forecast System, "NAM" stands for the North American Mesoscale model, and "HAFS" stands for the Hurricane Analysis and Forecasting System.

Name	Prototype	Microphysics	PBL	Surface layer	LSM
M0B0L0	*RRFS control*	*Thompson*	*MYNN-EDMF*	*MYNN*	*Noah*
M0B0L1		Thompson	MYNN-EDMF	MYNN	NoahMP
M0B0L2	*HRRR*	*Thompson*	*MYNN-EDMF*	*MYNN*	*RUC*
M0B1L0		Thompson	Shin–Hong	GFS	Noah
M0B2L0		Thompson	TKE-EDMF	GFS	Noah
M0B2L1	*Future GFS*	*Thompson*	*TKE-EDMF*	*GFS*	*NoahMP*
M1B0L0	*WoFS*	*NSSL*	*MYNN-EDMF*	*MYNN*	*Noah*
M1B0L1		NSSL	MYNN-EDMF	MYNN	NoahMP
M1B1L0		NSSL	Shin–Hong	GFS	Noah
M1B2L0		NSSL	TKE-EDMF	GFS	Noah
M2B0L0		Morrison–Gettelman	MYNN-EDMF	MYNN	Noah
M2B0L1		Morrison–Gettelman	MYNN-EDMF	MYNN	NoahMP
M2B1L0		Morrison–Gettelman	Shin–Hong	GFS	Noah
M2B2L0		Morrison–Gettelman	TKE-EDMF	GFS	Noah
M3B3L0	*NAM, HAFS*	*Ferrier–Aligo*	*K-EDMF*	*GFS*	*Noah*

cases from the full dataset (the green colored days in Fig. 1b). These forecasts use additional physics combinations beyond the five used in the HMT forecasts. We will call this the "expanded set of physics configurations" to distinguish it from the HMT forecasts that were run in real time for the HMT WWE. The expanded set selects PBL and LSM options to create more forecast pairs with only one physics parameterization difference at a time so that differences among the forecasts can be attributed to individual schemes rather than certain combinations only. For example, the M0B2L0 physics configuration is the same as the M0B0L0 physics configuration except that the TKE-EDMF PBL scheme is used instead of the MYNN scheme. The M0B0L1 physics configuration has the options of the proposed control member of the operational RRFS. In addition, the Morrison–Gettelman microphysics scheme (Morrison and Gettelman 2008), another fully two-moment microphysics scheme, and the Shin–Hong PBL scheme (Shin and Hong 2015), a scale-aware PBL scheme based on the widely used *K*-profile nonlocal YSU scheme, are added to allow for their examination of their relative performance.

All forecasts except for M3B3L0 use the NOAA Global Systems Laboratory (GSL) fork of FV3-LAM, checked out from https://github.com/NOAA-GSL/ufs-weather-model on 16 October 2020. The NSSL microphysics scheme used in the M1B0L0 physics configuration is merged with this version. M3B3L0 uses the Hurricane Analysis and Forecasts System (HAFS) community fork of FV3-LAM checked out from https://github.com/hafs-community/ufs-weather-model on 11 October 2020. The reason for the version discrepancy is that at the time, the Ferrier–Aligo scheme was not working in the GSL version. To the best of the authors' knowledge, these version differences do not make a substantial impact on the results.

In addition to the CAM forecasts, 6-hourly output from the operational GFSv15 forecast, which uses the FV3 dynamic core on a global 13-km grid and uses GFDL microphysics, K-EDMF PBL, GFS surface layer, and Noah LSM, is used as a point of comparison. The GFS and forecasts herein share the same initial conditions, and both use the FV3 dynamic core, so the grid spacing and physics are the main differences. GFS forecasts also allow for comparison of the full 84-h length of the forecasts herein. In contrast, the operational High-Resolution Rapid Refresh (HRRR) uses different IC/LBCs and a different dynamic core, and its forecasts only go to 48 h.

b. Evaluation methods

All evaluations are performed using the Model Evaluation Tools (MET) software package v9.1, specifically the grid_stat program, for grid-to-grid evaluations. Many different grids must be considered, and several methods are used for regridding depending on the scales involved. Re-gridding from coarse to fine scales ("downscaling") uses MET's budget method, which conserves mass in a grid box. Re-gridding from fine to coarse scales ("upscaling") uses MET's area-weighted mean method. Re-gridding between grids on the same scale uses MET's nearest-neighbor method. The analysis products used for verification cover the CONUS and nearby ocean regions, but we restrict our evaluations to the areas over or within 50 km of the CONUS (Fig. 1a) to avoid poorly observed areas far from land. Also, the CONUS is divided into five verification subdomains to allow regional verification (Fig. 1a).

Upper-air geopotential height, wind, temperature, and relative humidity forecasts at various levels are evaluated using the 6-hourly GFS 0.25° final analyses as truth. All model grids are interpolated to the GFS 0.25° grid using the area-weighted mean method. In addition, the verification fields are masked where pressure surfaces are below ground level to remove data that are not physically meaningful. In addition to the upper-air fields, surface fields such as 2-m temperature and dewpoint and 10-m wind are evaluated against the Unrestricted Mesoscale Analysis (URMA). The URMA is similar to the

Real Time Mesoscale Analysis (RTMA; De Pondeca et al. 2011) except it runs 6 h later on the same grid in order to use observations that arrive too late to be incorporated into the RTMA (Pondeca et al. 2015). The URMA is designed to have high fidelity to surface observations, so it is acceptable to use for forecast validation. The surface fields from all CAMs are interpolated to the 2.5-km URMA grid using the nearest-neighbor method. For the GFS, the budget method is used for interpolation. For both surface and upper-air fields, bias and root-mean-square error are evaluated.

Precipitation is compared to the NCEP Stage-IV QPE analysis (Nelson et al. 2016). The Stage-IV analysis is created by blending rainfall estimates from the WSR-88D radars with observations from rain gauges, followed by human corrections. All datasets are interpolated to the 4-km Stage-IV grid for comparisons. The convective-scale forecasts use the nearest-neighbor method for interpolation, and the operational GFS forecast uses the budget interpolation method as with the URMA grid. In addition to QPF, snowfall is evaluated against the National Operational Hydrologic Remote Sensing Center (NOHRSC) snowfall analysis version 2, which is a blend of Stage-IV QPE, HRRR, RAP analyses, and snowfall reports. For the forecasts, snowfall is taken from the snow and cloud ice mass reaching the surface in the microphysics scheme and the accumulation is converted to a snow depth using a 10:1 snow-liquid depth ratio. At locations where the precipitation type algorithm does not diagnose snow, snowfall is set to zero. For all forecasts except M3B3L0, the HRRR explicit precipitation type algorithm (Benjamin et al. 2016) is used. The HRRR precipitation type algorithm cannot be applied directly to the M3B3L0 forecast because the Ferrier–Aligo microphysics does not have separate snow and graupel categories, as required by the algorithm. Therefore, a column-temperature-based algorithm (Manikin 2005) is used for this forecast as a fallback. In addition, accumulated snowfall is not available in the archived GFS forecasts; however, snow depth is available. Therefore, the change in snow depth over the accumulation interval is considered for the GFS instead of accumulated snowfall, and instances where the snow depth change is negative are set to zero. The implications of this are discussed in section 5. All CAM snowfall forecasts are interpolated to the NOHRSC grid using the nearest-neighbor method, while re-gridding the operational GFS snowfall forecast uses the budget method.

For each forecast lead time and case, a 2×2 contingency table is constructed from the model and analysis snowfall and QPF using several thresholds. A threshold of 1 mm is used to distinguish precipitation from no precipitation, and thresholds of 25 and 50 mm are used to identify heavier precipitation. A threshold of 2.5 cm is used to distinguish snow from no snow. From these contingency tables, equitable threat score (ETS) and frequency bias are computed as in Mason (2003). To compute aggregated statistics, the elements of the contingency tables for all cases are summed, and then ETS and frequency bias are computed from the aggregated contingency table. In previous work, ETS has been found to reward forecasts with a higher bias for a given displacement error for rare events (Baldwin and Kain 2006). However, because of the low precipitation threshold, long accumulation period, and the usual stratiform nature of winter precipitation, and because this dataset contains a disproportionately large number of cases with widespread precipitation coverage, the event frequency of QPF ≥ 1 mm for this dataset is 19.4%. This is closer to Baldwin and Kain's (2006) "common event" frequency level (28%) at which the apparent reward by ETS for a high bias is greatly reduced. Also, despite the advantages of neighborhood-based contingency tables, for simplicity, we do not use a neighborhood when computing contingency tables herein. The results are qualitatively the same with and without a neighborhood (not shown), likely because the 24-h accumulations have a relatively large spatial coverage and already implicitly allow for temporal error (see sections 4 and 5).

Additionally, we use the resampling method of Hamill (1999) for statistical significance testing of verification scores. Essentially, we propose the null hypothesis that $Q_{M1} - Q_{M2} = 0$, where Q is a contingency table statistic (e.g., frequency bias) and M1 and M2 denote forecasts from two different physics configurations. To test this hypothesis, we create a 1000-sample null difference distribution by resampling the contingency tables as in Hamill (1999), and the null hypothesis is rejected at significance level p if the actual $Q_{M1} - Q_{M2} = 0$ falls outside the $p/2$ to $1 - (p/2)$ percentile range of the null difference distribution. For these experiments, $p = 0.01$, i.e., 1%.

3. Surface field verification

The 2-m temperature bias in most of the HMT forecasts (Fig. 2a) across the CONUS displays a strong diurnal cycle in all forecasts: generally, a cool bias, with the bias being smaller in magnitude overnight and in the early part of the day (0600–1800 UTC) and larger in magnitude during the day and into the evening. The exception is the M3B3L0 forecast which, along with the GFS, generally has a warm bias overnight and a cool bias during the day. The warm bias overnight is particularly pronounced over the southern plains (not shown). Both the M3B3L0 and GFS forecasts use the K-EDMF PBL scheme, suggesting that this scheme tends to produce a warm bias during the overnight hours. The reasons for this are being investigated in depth in a separate study. In addition to M3B3L0 and GFS generally having comparable or smaller levels of bias than other forecasts, they also generally exhibit lower RMSE (Fig. 2b) than all the other forecasts. Furthermore, the M0B0L0 and M1B0L0 forecasts, which differ only in the use of Thompson versus NSSL microphysics, broadly display very little difference either in bias or RMSE, despite the two forecasts generally being statistically significantly different than each other. On one hand, this is not unexpected, as 2-m temperature is most directly impacted by PBL, surface-layer, and land surface processes and less so by the microphysics. However, the microphysics can impact the surface temperature through differences in radiation due to clouds, differences in snow cover, or differences in latent heating and cooling. These results suggest that if these differences exist between Thompson and NSSL, they are either not large enough or not systematic

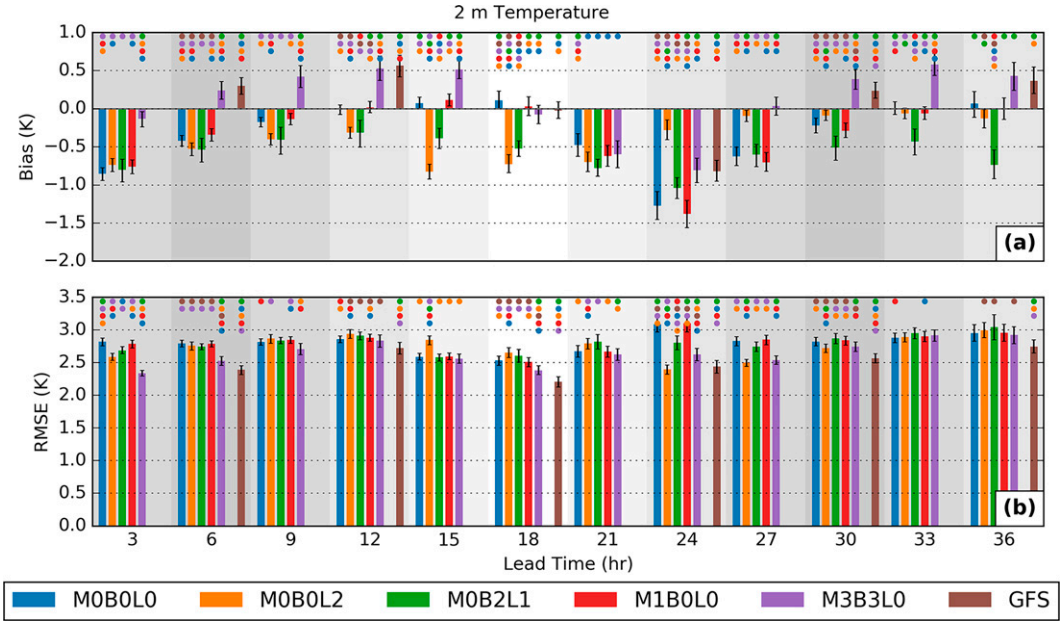

FIG. 2. (a) Bias and (b) RMSE for 2-m temperature for the HMT forecasts over the CONUS as a function of lead time for all cases in the winter season. All forecasts are initialized at 0000 UTC. The background shading is darkest at 0600 UTC (approximately midnight local time over the CONUS) and lightest at 1800 UTC (approximately noon local time over the CONUS). The error bars give the 2.5–97.5 percentile of the quantity bootstrap resampled 1000 times over all cases. Colored dots above each bar indicate other forecasts that are statistically significantly ($p = 0.01$) different than that forecast.

enough (e.g., overall cloud cover may be small) in the cool season to noticeably impact the bulk seasonal statistics.

For 2-m dewpoint temperature bias (Fig. 3a), all forecasts generally have a moist bias. The exception is the M0B2L1 forecast that uses the NoahMP LSM. M0B2L1 displays a prominent negative bias starting at 15 h into the forecast (largest at 18–21 h, during the afternoon hours over the CONUS). The forecast with the largest moist bias is the M0B0L2 forecast, again peaking during the afternoon and early evening. The M3B3L0 forecast and the operational GFS are closest to being unbiased and are again among the lowest in terms of RMSE (Fig. 3b). Sharing the same PBL, surface layer physics, and LSM is the likely reason for the similarity between these two forecasts.

To diagnose the reason for the differences in the 2-m dewpoint forecasts, we examine the total daily latent heat flux (Fig. 4). Latent heat flux in the control forecast (Fig. 4a) is generally higher over bodies of water, with enhanced latent heat flux over the Gulf Stream. M0B0L2 (Fig. 4b) exhibits a similar pattern; the calculation in the GFS NSST scheme for latent heat flux over water uses the surface exchange coefficient for heat computed by the surface layer scheme, and both M0B0L0 and M0B0L2 use the MYNN surface layer scheme. Additionally, latent heat flux is generally higher in M0B0L2 than in M0B0L0 over areas of the southeastern United States and mountain west, while areas of Texas and northern Mexico have lower latent heat flux. These differences are attributable directly to the LSM used, as over land, latent heat flux is calculated within the LSM. The latent heat flux in M0B2L1 (Fig. 4c), which uses NoahMP, is nearly universally lower over land and higher over water than in M0B0L0. M3B3L0 (Fig. 4d), despite having the same LSM (Noah) as the control, exhibits some notable differences. First, latent heat flux over water is higher in M3B3L0 than M0B0L0 (though still less than M0B2L1), and latent heat flux over the mountains of the northwest United States and southwestern Canada is also higher in M3B3L0 than in M0B0L0. Additionally, artifacts are visible in the southeastern United States (Fig. 4d) caused by the relatively coarse resolution of the vegetation type data used for the M3B3L0 forecasts. The GFS surface layer used in M3B3L0 is more sensitive to changes in vegetation type than the MYNN surface layer, so the artifacts are particularly obvious. These artifacts do not have a substantial negative impact on the overall forecast performance based on sensitivity experiments performed (not shown).

The higher latent heat flux computed by the RUC LSM over land helps explain the larger positive bias of 2-m dewpoint in M0B0L2 versus M0B0L0. Additionally, the lower latent heat flux computed by the NoahMP LSM over CONUS in M0B2L1 helps explain its low bias in the bulk dewpoint statistics (Fig. 3a), with the increases in latent heat flux caused by using the GFS surface layer scheme in M0B2L1 versus the MYNN surface layer scheme in M0B0L0 over water falling largely outside the verification region, which extends only 50 km from the coastline. These results are supported by spatial patterns in the dewpoint bias, which generally mirror those in the latent heat flux (not shown).

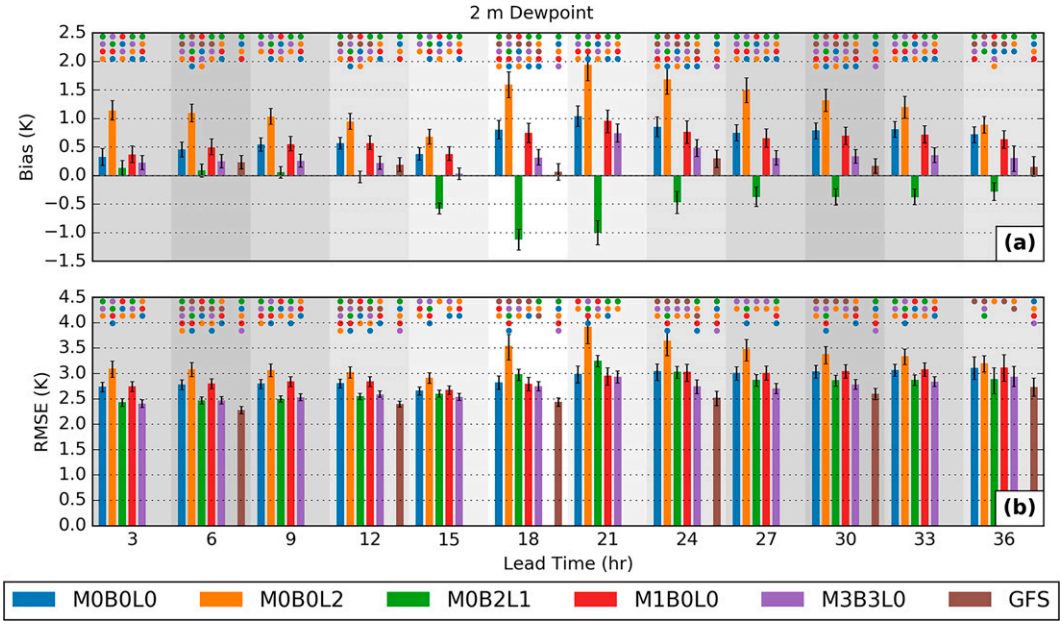

FIG. 3. As in Fig. 2, but for 2-m dewpoint temperature.

4. Precipitation evaluation

a. *Bulk statistics*

Frequency bias for 24-h precipitation at a threshold of 1 mm (rain or snow liquid equivalent) from the HMT forecasts over the full CONUS domain (Fig. 5) is generally small, but positive (i.e., greater than 1), indicating that all forecasts generally over predict total precipitation coverage. M1B0L0 has the lowest frequency bias of all the forecasts at all lead times, followed by M0B0L0 (Fig. 5a). Both M0B0L0 and M1B0L0 outperform the operational GFS forecast, and M0B2L1, which has the largest positive bias. ETS (Fig. 5b) slightly decreases over the forecast period, and forecasts are generally more similar to each other in ETS than in frequency bias, though some differences exist. In particular, M0B2L1 has the lowest ETS. Also, although M1B0L0 has significantly lower bias than all other forecasts, its ETS is generally not significantly different.

For the higher threshold of 50 mm in 24 h (Fig. 6), the positive bias in precipitation coverage in the CAM forecasts is reduced, particularly at the 60- and 84-h forecasts, but still present at 36 h, suggesting the positive frequency bias occurs primarily, but not exclusively, with light precipitation. In addition, the operational GFS forecast has large negative biases (i.e., less than 1) at the 50 mm threshold, which also drive down its ETS compared to several of the other forecasts (Fig. 6b). The low bias with heavy precipitation is consistent with Jiang et al. (2017, see their Fig. 4a), and Ganai et al. (2021, see their Fig. 1) also find that the GFS has low precipitation biases with heavier (≥15 mm) 24-h precipitation. Also, Zhu et al. (2018) find that 24-h warm-season precipitation has a low bias for thresholds ≥ 50 mm in 24 h in four global operational forecasting systems, while 4-km forecasts based on the WRF Model have high biases.

On the subset of cases used for the expanded set of physics configurations (Fig. 7), many of the same behaviors with precipitation are present as in the full set of cases used for the HMT forecasts. Overall, the precipitation bias is strongly clustered by microphysics scheme (the bars in Fig. 7 are grouped by microphysics scheme), except for M0B2L0 and M0B2L1. The clustering by microphysics scheme is consistent with the findings of Zhang et al. (2019). The M1B0L0, M1B0L1, M1B1L0, and M1B2L0 physics configurations, which have the NSSL MP scheme in common, generally perform well—they are generally the closest to being unbiased (i.e., a frequency bias of 1) and among the forecasts with the highest ETS. The M2B0L0, M2B0L1, M2B1L0, and M2B2L0 configurations, which have the Morrison–Gettelman MP scheme in common, are generally negatively biased (in contrast with the positive bias of the other forecasts) and have lower ETS than many of the Thompson and NSSL MP forecasts.

In addition, we can contrast M0B0L0, M0B2L0, M1B0L0, and M2B0L0 (which have the Noah LSM in common) to M0B0L1, M0B2L1, M1B0L1, and M2B0L1 (which have the NoahMP LSM in common). These forecasts are chosen for comparison because they are pairs of physics configurations that only differ in the LSM (e.g., M0B0L0 uses the Noah LSM and M0B0L1 uses the NoahMP LSM, but the two are otherwise identical). The NoahMP LSM forecasts have smaller (i.e., closer to 1) bias, except for M2B0L1, which has a larger negative bias. Also, the NoahMP forecasts generally have slightly higher ETSs than the Noah forecasts for all pairs of forecasts at all lead times, suggesting that changes in NoahMP from Noah act to reduce precipitation in cool season CAM forecasts.

The expanded set of physics configurations allows us to isolate which parameterizations are responsible for the behaviors noted in the set of five HMT forecasts. For 24-h QPF, the patterns in skill scores in HMT forecasts (Fig. 5) generally hold

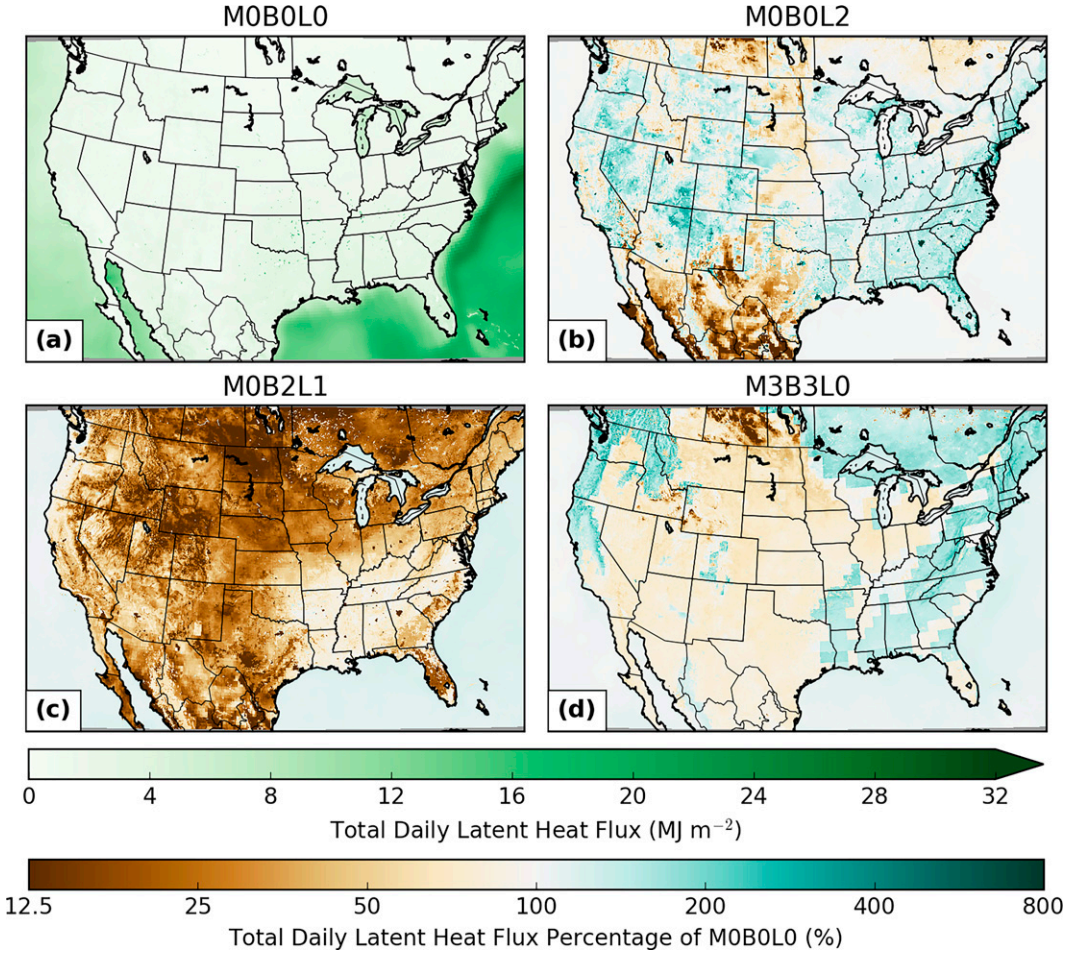

FIG. 4. Total daily latent heat flux for the 24-h period ending at F36. Shown here is the mean of 30 cases for (a) M0B0L0. Additionally, (b) M0B0L2, (c) M0B2L1, and (d) M3B3L0 are shown as percentages of M0B0L0.

when run on the subset of 8 cases used for the expanded set of physics configurations. For example, on this subset of cases, M0B2L1 has the highest frequency bias and lowest ETS, and M1B0L0 has the lowest bias (Fig. 7a). These are the same behaviors displayed on the full set of 35 cases (Fig. 5a). This gives us confidence to draw conclusions from the expanded set of physics configurations despite the comparatively smaller sample size. In this case, the similarity between M0B2L0 and M0B2L1 noted above suggests that the TKE-EDMF PBL scheme in combination with Thompson microphysics, rather than the NoahMP LSM, is primarily responsible for the poor performance of M0B2L1. This is confirmed by M0B0L1, which uses the MYNN PBL scheme instead of TKE-EDMF PBL, and the scores of M0B0L1 and M0B0L0 are much closer. In contrast to M0B2L1, M0B1L0, using the Shin–Hong PBL scheme, is comparable to M0B0L0 for bias and ETS for all lead times.

To better understand the behaviors of the different forecasts, we look at the spatial distribution of hits, misses, false alarms, and correct nulls (Fig. 8) for forecasts of ≥1 mm of precipitation in 24 h. The color in this figure is based on a ternary plot, which considers the values of three variables that sum to 1. The contingency table components sum to 1 everywhere in the domain, so each point in the domain is colored based on its position in the ternary plot, based on the key at the bottom of Fig. 8. The hits and correct nulls are summed for one axis of the ternary plot in order to reduce the number of variables to three and leave misses and false alarms with their own axes. Thus, areas in orange on Fig. 8 have more cases with false alarms, areas in blue have more cases with misses, and areas in dark gray have an even mix. The areas in white have a perfect forecast. This has the advantage over a frequency bias plot in that it distinguishes between areas with few misses and false alarms and areas where misses and false alarms cancel each other.

Over the CONUS as a whole, false alarms slightly outnumber misses (not shown), consistent with all forecasts having a positive bias in Fig. 5a. The false alarms are not concentrated in just a few cases, suggesting over prediction of precipitation coverage is a systemic feature of these forecasts. Additionally, prominent areas of much higher false alarms appear over the

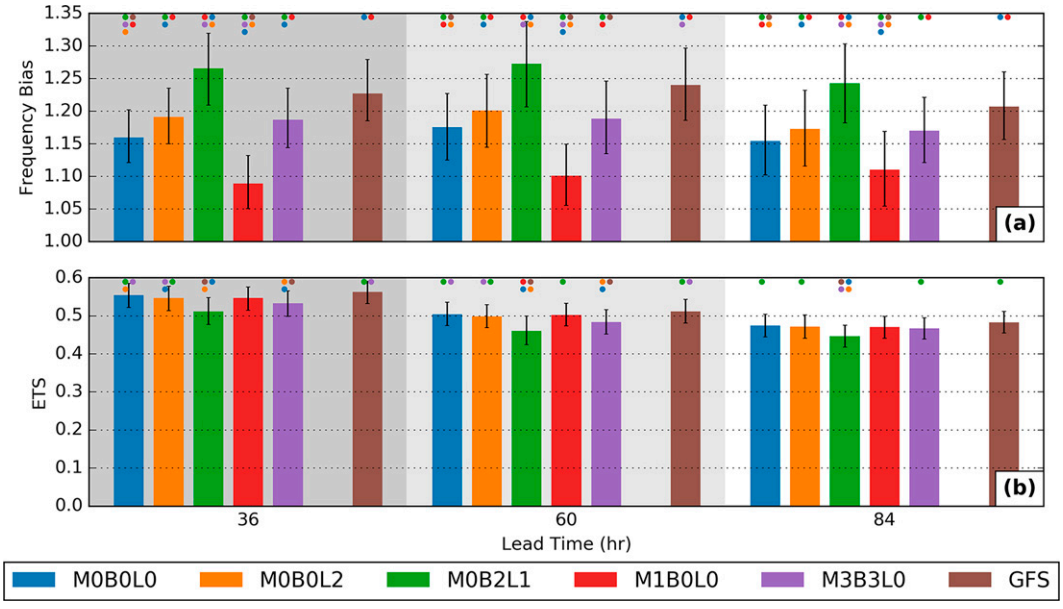

FIG. 5. (a) Frequency bias and (b) equitable threat score for 24-h QPF over CONUS as a function of lead time for all cases in the winter season. A QPF threshold of 1 mm is used. The colored dots and error bars are as in Fig. 2, but the cases are bootstrap resampled 10 000 times to create the error bars.

eastern Great Lakes and northern Appalachians (Fig. 8) in all forecasts. M0B2L1 (Fig. 8e) has many more false alarms than other physics configurations over Lakes Superior, Huron, and to a lesser extent Michigan. M1B0L0 (Fig. 9b) is unique in having a relatively large number of misses in the Gulf Coast states, which accounts for the lower CONUS-wide bias in those forecasts. The operational GFS and M0B2L1 (Figs. 9f and 9e, respectively) also have a much higher proportion of false alarms offshore in south Florida than other forecasts. Because the positive bias in precipitation coverage noted in Fig. 7 is driven by over prediction of QPF in these regions, and this positive bias only occurs in M0B2L0 and M0B2L1, this implies that the higher false alarm count over south Florida and the Great Lakes seen in M0B2L1 in Fig. 8 is the result of the TKE-EDMF PBL scheme in combination with Thompson microphysics. Similarly, the NSSL microphysics scheme is responsible for the higher miss count along the gulf coast.

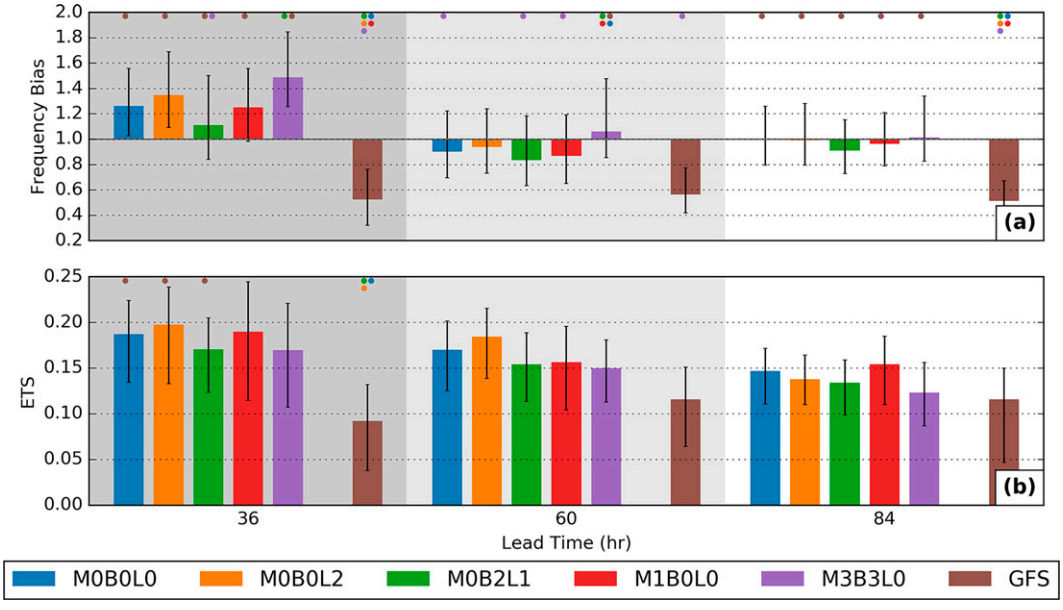

FIG. 6. As in Fig. 5, but with a 24-h QPF threshold of 50 mm.

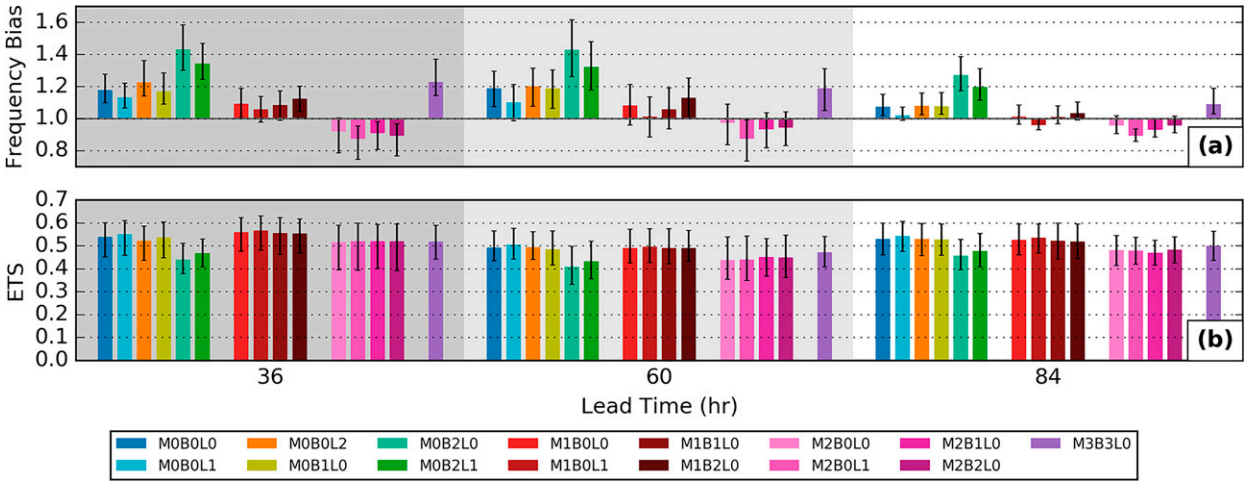

FIG. 7. As in Fig. 5, but for the expanded set of 15 physics configurations (8 cases). The statistical significance markers are not shown.

b. NSSL microphysics behavior

The six cases with the most prominent miss count in M1B0L0 are 15 December, 29 December, 25 January, 30 January, 4 February, and 2 March. Many of these are associated with weak convection in low-CAPE environments (not shown). From these cases, 30 January is chosen as a representative case for further examination. The contingency table components for the 36-h forecast of 24-h QPF for this case for all forecasts are shown in Fig. 10. M1B0L0 (Fig. 10b) has by far the most misses (and few hits and false alarms) in Louisiana and southern Mississippi and Alabama. By contrast, most other forecasts, including the operational GFS, nearly exclusively have hits and false alarms in this region.

One reason for this can be found in comparing profiles from M0B0L0 and M1B0L0 over Louisiana and southern Mississippi (Fig. 11). As expected from the differences in precipitation coverage, the 98th percentile of rain mixing ratio in M0B0L0 is much higher at the surface than M1B0L0 (Figs. 11c,g), indicating more precipitation reaching the surface in M0B0L0. In both forecasts, all precipitation is created as rain, which is consistent with the moist layer in both forecasts being entirely below the freezing level (Figs. 11a,e). Consistent with the differences in rain mixing ratios, the mean mass diameter (MMD) of the raindrop size distributions in M0B0L0 are much larger than in M1B0L0 (Figs. 11d,h), indicating that the Thompson scheme produces larger drops than the NSSL scheme. Furthermore, the 98th percentile of rain mixing ratio in the NSSL scheme decreases toward the surface, while the MMD increases toward the surface, which is consistent with the evaporation of the smaller drops in the size distributions due to the smaller drops having a larger surface area to volume ratio (Dawson et al. 2010). This suggests evaporation of the smaller drops plays a major role in the QPF differences between the Thompson and NSSL schemes. The differences in drop sizes arise despite few differences in the vertical velocity distribution (Figs. 11b,f) and the environmental temperature and moisture profiles (Figs. 11a,e), suggesting that they are inherent to the formulations of the respective schemes.

c. The TKE-EDMF PBL scheme behavior

Next, we examine the causes of the higher bias in M0B2L1 for precipitation near the Great Lakes. The localized nature of the precipitation enhancements near bodies of water suggests enhanced latent heat flux as a cause. Indeed, the total daily latent heat flux is higher in M0B2L1 than in M0B0L0 (Fig. 4c). Supporting this, the 850-hPa relative humidity bias over the northeast verification region (Fig. 12b) is positive for M0B2L0 and M0B2L1 in the expanded set of physics configurations even starting at the 6-h forecast. These two physics configurations have the TKE-EDMF PBL scheme and Thompson microphysics in common, again suggesting the combination of these two schemes is responsible for the high precipitation bias. M1B2L0 and M2B2L0, also using the TKE-EDMF PBL scheme, but with the NSSL and Morrison–Gettelman microphysics schemes, respectively, rather than the Thompson microphysics scheme, also show a positive bias in 850-hPa relative humidity, though not to the extent as in the runs with Thompson microphysics. M0B2L0 and M0B2L1 have the lowest 850-hPa relative humidity RMSE (not shown), consistent with the higher biases, at least in the first 36 h of the forecast. The primary reason for the positive relative humidity biases appears to be a positive bias in dewpoint at 850 hPa (Fig. 12c). Temperature bias at 850 hPa (Fig. 12a) is negative, which also contributes to a positive relative humidity bias; however, the magnitude of the temperature bias is smaller than that of the dewpoint bias.

M3B3L0 also has enhanced latent heat flux over the Great Lakes compared to M0B0L0 (Fig. 4d), though to a lesser degree than M0B2L1. M3B3L0 and M0B2L1 share the GFS surface layer scheme which, as mentioned above, has a strong control over the latent heat flux calculation over water via the surface exchange coefficient for sensible and latent heat. However, M3B3L0 has a low bias in 850-hPa relative humidity. The differences between M3B3L0 and M0B2L1 must be related to the way their respective PBL schemes, i.e., the TKE-EDMF and K-EDMF PBL schemes handle the vertical mixing of near-surface moisture. Thus,

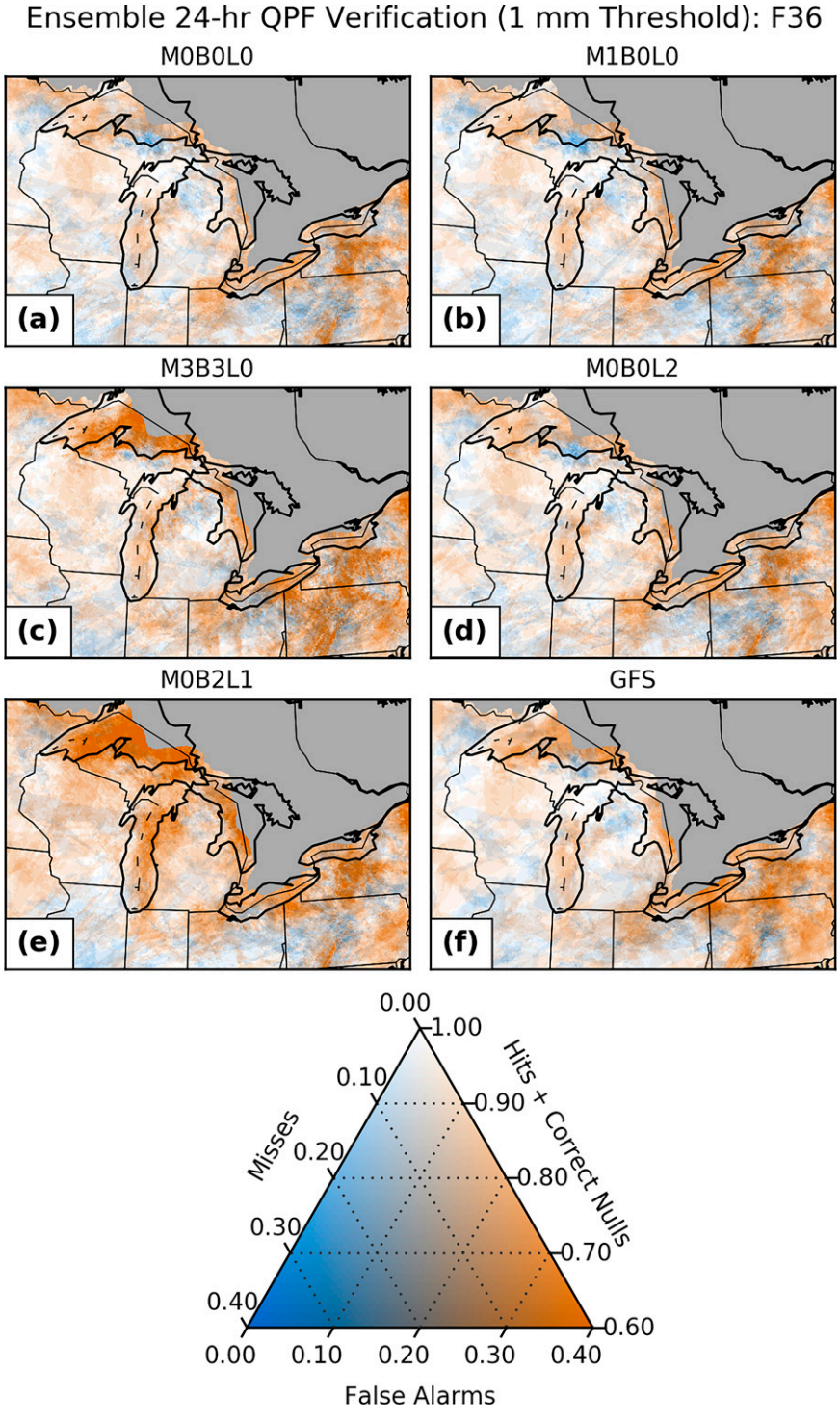

FIG. 8. Proportion of misses and false alarms in 24-h QPF over the Great Lakes, aggregated over all cases at a 36-h lead time using a threshold of 1 mm for (a) M0B0L0, (b) M1B0L0, (c) M3B3L0, (d) M0B0L2, (e) M0B2L1, and (f) GFS.

we conclude that enhanced latent heat fluxes over the Great Lakes, primarily caused by the GFS surface layer scheme, lead to an over prediction of precipitation over the Great Lakes through vertical mixing by the TKE-EDMF PBL and realization of precipitation by Thompson microphysics. This same basic process is also likely responsible for the over prediction of precipitation offshore in south Florida (not shown).

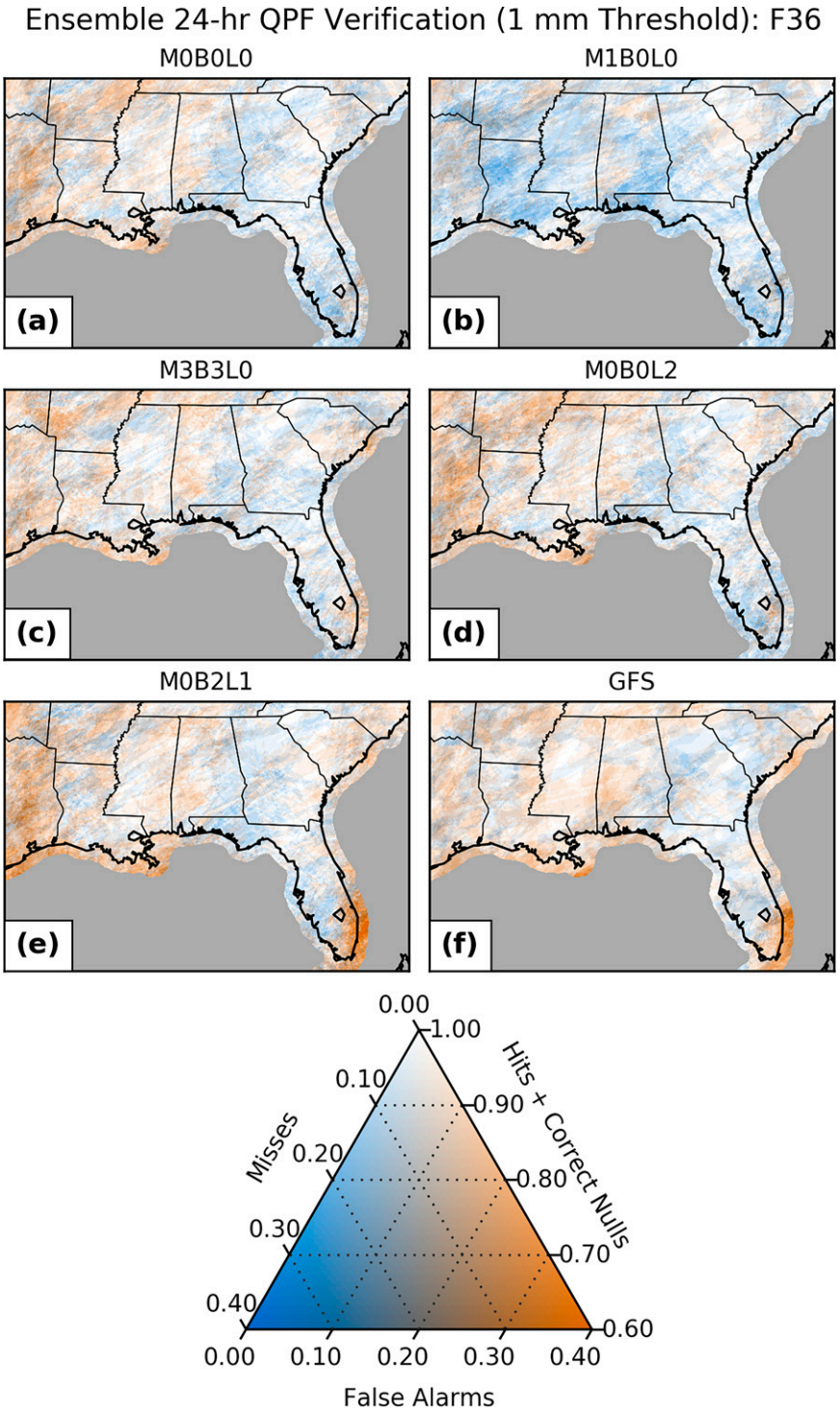

FIG. 9. As in Fig. 8, but for the southeast United States.

5. Snowfall evaluation

The 24-h snowfall from the HMT forecasts (Fig. 13) shares many of the behaviors of QPF. Many forecasts have a positive bias, with M0B2L1 having the highest bias at all lead times. ETS for snowfall generally decreases with increasing lead time. In contrast to the QPF, M3B3L0 and GFS are nearly unbiased for snowfall, and the GFS has an ETS that is not significantly different that of M0B0L0. However, M3B3L0 has a

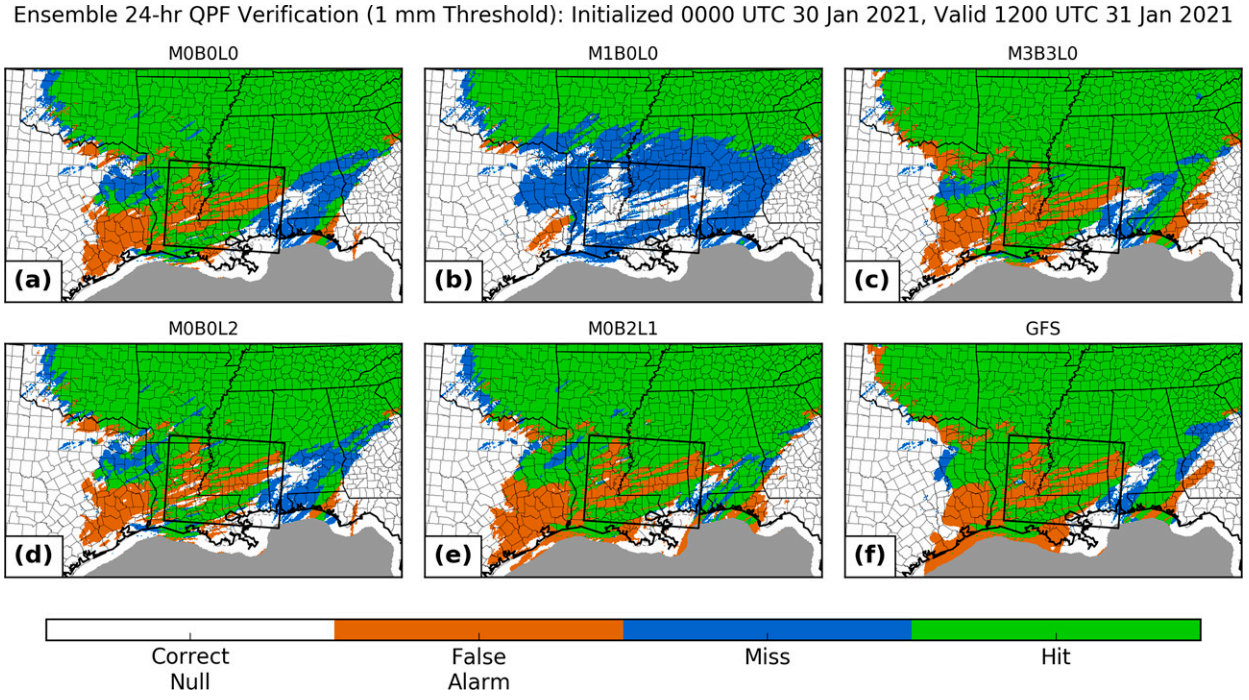

FIG. 10. Components of the contingency table for 24-h QPF for (a) M0B0L0, (b) M1B0L0, (c) M3B3L0, (d) M0B0L2, (e) M0B2L1, and (f) GFS. Green areas indicate hits, blue indicates misses, orange indicates false alarms, and white indicates correct negatives. Shown is the 36-h forecast from the 30 January case. The black box indicates the area over which the profiles are taken in Fig. 11.

significantly lower ETS than all the other forecasts and the GFS, despite being nearly unbiased.

As with QPF, the false alarms in 24-h snowfall at forecast hour 36 (i.e., day 1) slightly outnumber the misses, which is consistent with most forecasts having an overall positive bias in snowfall. M0B2L1 also shows larger numbers of false alarms than other forecasts near the Great Lakes (not shown), consistent with the behavior in QPF. However, there are some differences between

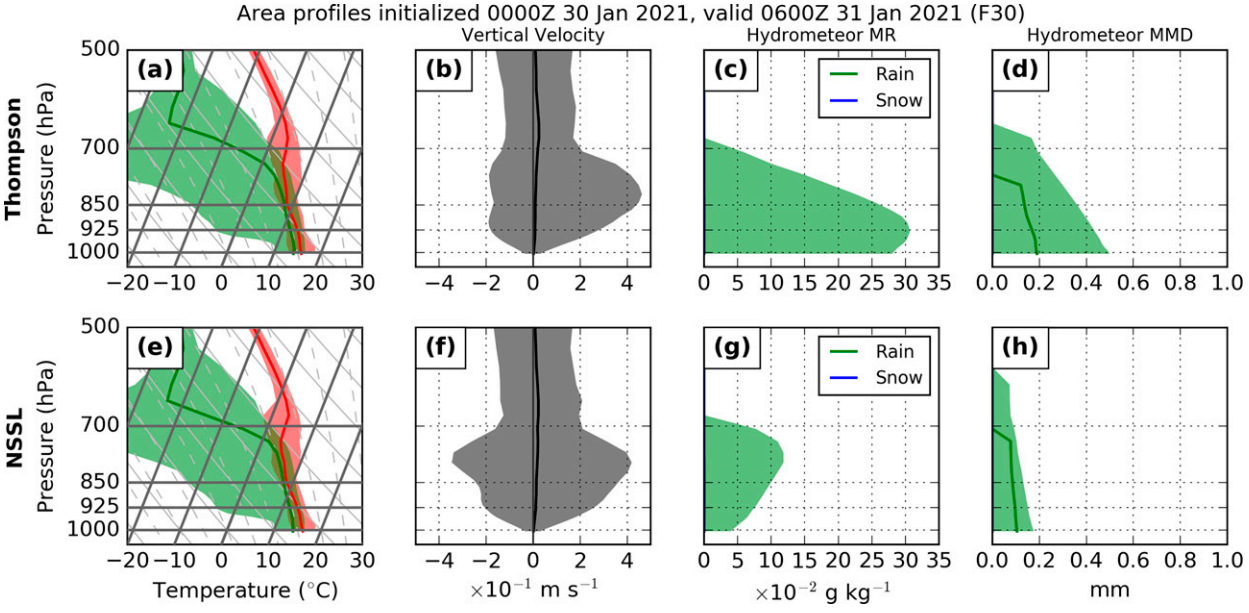

FIG. 11. Profiles to 500 hPa over all columns in the black box in Fig. 10. (a),(e) Skew T–log$p$ diagram; (b),(f) vertical velocity; (c),(g) hydrometeor mixing ratios (MR); and (d),(h) hydrometeor mean mass diameter (MMD). (top) M0B0L0 and (bottom) M1B0L0. The thick solid line is the median of all model profiles in the selected region, and the shaded region represents the 2nd–98th percentile range of all columns. The coverage of nonzero hydrometeor MR is less than 50% at all levels, therefore the median hydrometeor MR is 0 at all levels.

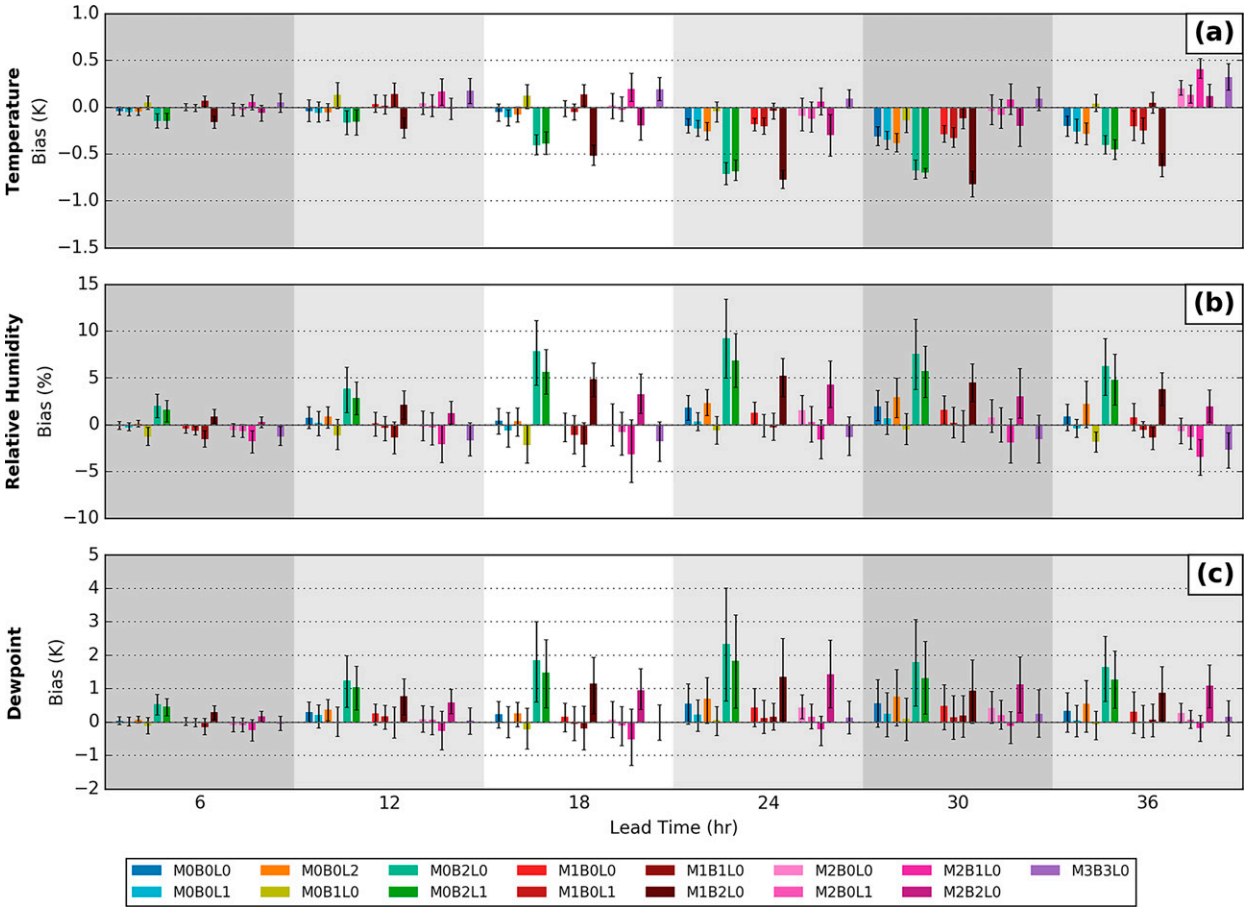

FIG. 12. As in Fig. 7, but for (a) 850-hPa temperature, (b) relative humidity, and (c) dewpoint from the northeast verification region for the first 36 h of the forecast.

the precipitation and snowfall scores. M3B3L0 and the GFS have more misses than false alarms near the Great Lakes, and in the southern Rocky Mountains (Fig. 14), which are not present in the precipitation scores.

Six cases in which the GFS missed snow in the southwest United States are prominent: 10 December, 25 January, 29 January, 16 February, 11 March, and 12 March. The 10 December and 12 March cases are selected as representative cases for further analysis (Fig. 15). One confounding factor with the GFS is the necessity of using of snow depth change to compare to an analysis of snowfall. This is because snow compacts over time, resulting in a decrease in snow depth, and in marginal temperatures (between 0° and ~3°C), snow may melt on contact with the ground, which results in less snow accumulating than the total amount of snow that falls. And many of the areas where the GFS misses 24-h snowfall accumulations of 2.5 cm for these cases fall into regions of marginal temperatures (Figs. 15a,b). In addition, many of the 24-h snowfall misses are also associated with positive temperature errors in the GFS. Some of this is likely due to the coarseness of the GFS terrain; some terrain features are obvious in southern Utah in Fig. 15b. Nevertheless, some error is probably due to the K-EDMF PBL scheme used. The M3B3L0 physics configuration, which also uses the K-EDMF PBL scheme, also has many of its 24-h snowfall misses in the southwest United States associated with positive temperature errors (Figs. 15c,d), despite using a finer terrain representation. Additionally, many of the misses in M3B3L0 occur in similar areas as those in the GFS (see e.g., the Four Corners region in Fig. 15a and Fig. 15c and southwest Utah in Fig. 15b and Fig. 15d). This further suggests the K-EDMF PBL scheme is responsible for some of the misses due to the forecast near surface temperature being too high. In contrast, the bias in 2-m temperature in the mountain west region (not shown) is negative for M3B3L0 at the 24-h forecast. There are a few regions where the GFS 24-h snowfall misses are associated with negative temperature errors (see Fig. 15a in Colorado, for example); however, those locations are below 0°C in the GFS forecast, making it quite unlikely that marginal temperatures and associated melting would be responsible for the miss. While Fig. 15 includes only 24-h forecast temperature errors, other forecast times during the day-1 period are qualitatively the same (not shown) with respect to temperature errors.

6. Summary and conclusions

This paper presents the results of evaluations of ~3-km grid spacing CAM forecasts based on the limited-area FV3 model

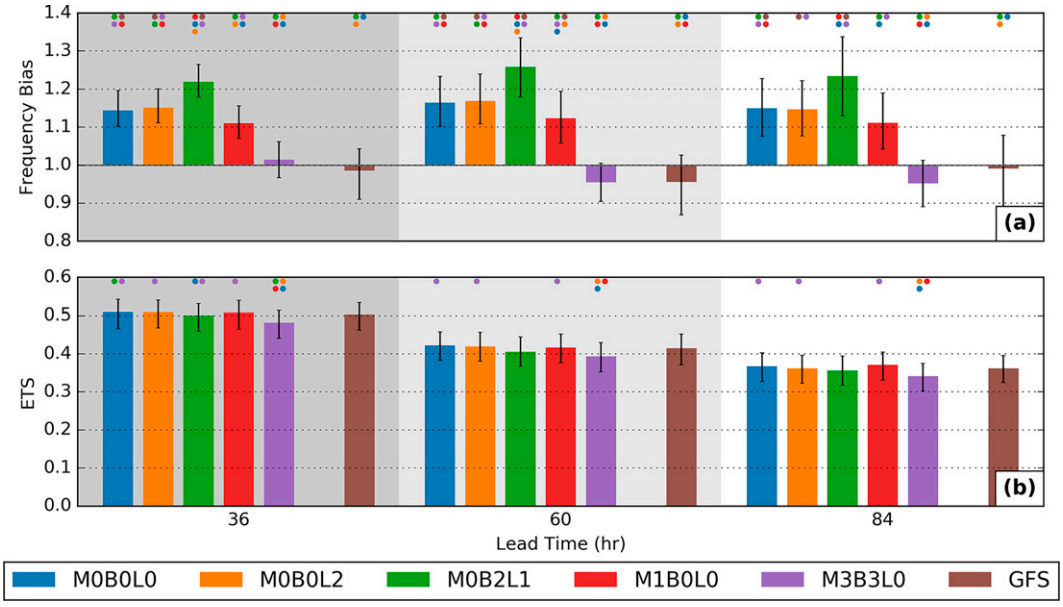

FIG. 13. As in Fig. 5, but for 24-h snowfall with a threshold of 2.5 cm.

during the 2020–21 winter season. The forecasts are grouped into two sets: a set of 5 physics configurations run on 35 cases during the 11th HMT WWE and an expanded set of 15 physics configurations run on a subset of eight cases from the HMT WWE. All forecasts were initialized at 0000 UTC on their respective dates and use NCEP GFS initial and lateral boundary conditions. Both sets of forecasts use varying combinations of microphysics (Thompson, NSSL, Morrison–Gettelman, and Ferrier–Aligo), PBL (MYNN, TKE-EDMF, K-EDMF, and Shin–Hong), surface layer schemes (MYNN PBL is used along with its surface layer scheme, and the other three PBL schemes are used with the GFS surface layer scheme) and land surface models (Noah, NoahMP, and RUC) to isolate the effect of individual schemes. All CAM forecasts use the GFS NSST scheme, which handles flux computations over water; flux computations over land are computed in the LSM. The CAM forecasts are compared to operational GFSv16 forecasts as a baseline. All forecasts are evaluated against several analysis products. Surface fields were evaluated against the URMA, upper-air fields were compared to GFS final analyses, precipitation forecasts were compared to the NCEP Stage-IV precipitation analysis, and snowfall forecasts were compared to the NOHRSC snowfall analysis version 2. For the surface and upper-air fields, bias and RMSE are used as metrics, and for precipitation and snowfall, frequency bias and ETS are used as metrics.

In the surface field evaluations, the K-EDMF PBL scheme is found to have a warm bias overnight. Additionally, there is little difference between the Thompson and NSSL microphysics schemes in terms of surface fields. The NoahMP LSM produces a pronounced low dewpoint bias during the afternoon, partially because of reduced latent heat flux compared to other forecasts. Furthermore, the RUC LSM is found to have a higher latent heat flux than the Noah LSM over land. The latent heat flux over water from forecasts using the GFS surface layer scheme is higher than that from forecasts using the MYNN surface layer scheme, likely a result of larger surface exchange coefficient for heat in the GFS surface layer scheme.

The precipitation evaluations show that all forecasts have a high bias on precipitation coverage for the 1 mm day^{-1} threshold. For the 50 mm day^{-1} threshold, the operational GFS has a low bias in coverage, while the CAM forecasts are generally high-biased or nearly unbiased. At the 1 mm day^{-1} threshold, the HMT forecasts using the NSSL microphysics scheme have the lowest positive frequency bias within the entire CONUS domain, but investigation shows that the NSSL microphysics under predicts precipitation in the Gulf Coast states. The NSSL microphysics scheme is found to produce smaller raindrops than the Thompson scheme; the smaller drops evaporate more readily, leading to less precipitation reaching the surface. The Morrison–Gettelman microphysics scheme is found to underforecast precipitation, and it also has reduced ETS compared to the Thompson and NSSL schemes. Also, the TKE-EDMF PBL scheme in combination with Thompson microphysics is associated with a high bias in precipitation over the Great Lakes and over the waters off the coast of Florida. Analyses attribute this to larger latent heat flux over water in these regions, leading to higher low-level relative humidity and more precipitation.

The snowfall evaluations show many of the same behaviors as the precipitation forecasts. Many of the CAM forecasts have a positive bias in coverage of 24-h snowfall $\geq$ 2.5 cm at all lead times. Additionally, the high precipitation bias in apparent in the TKE-EDMF PBL scheme when combined with Thompson microphysics is apparent in the snowfall forecasts. However, one difference is that the operational GFS and M3B3L0 display a low bias in snowfall in the southern Rocky Mountains, due to positive temperature errors produced by the K-EDMF PBL scheme.

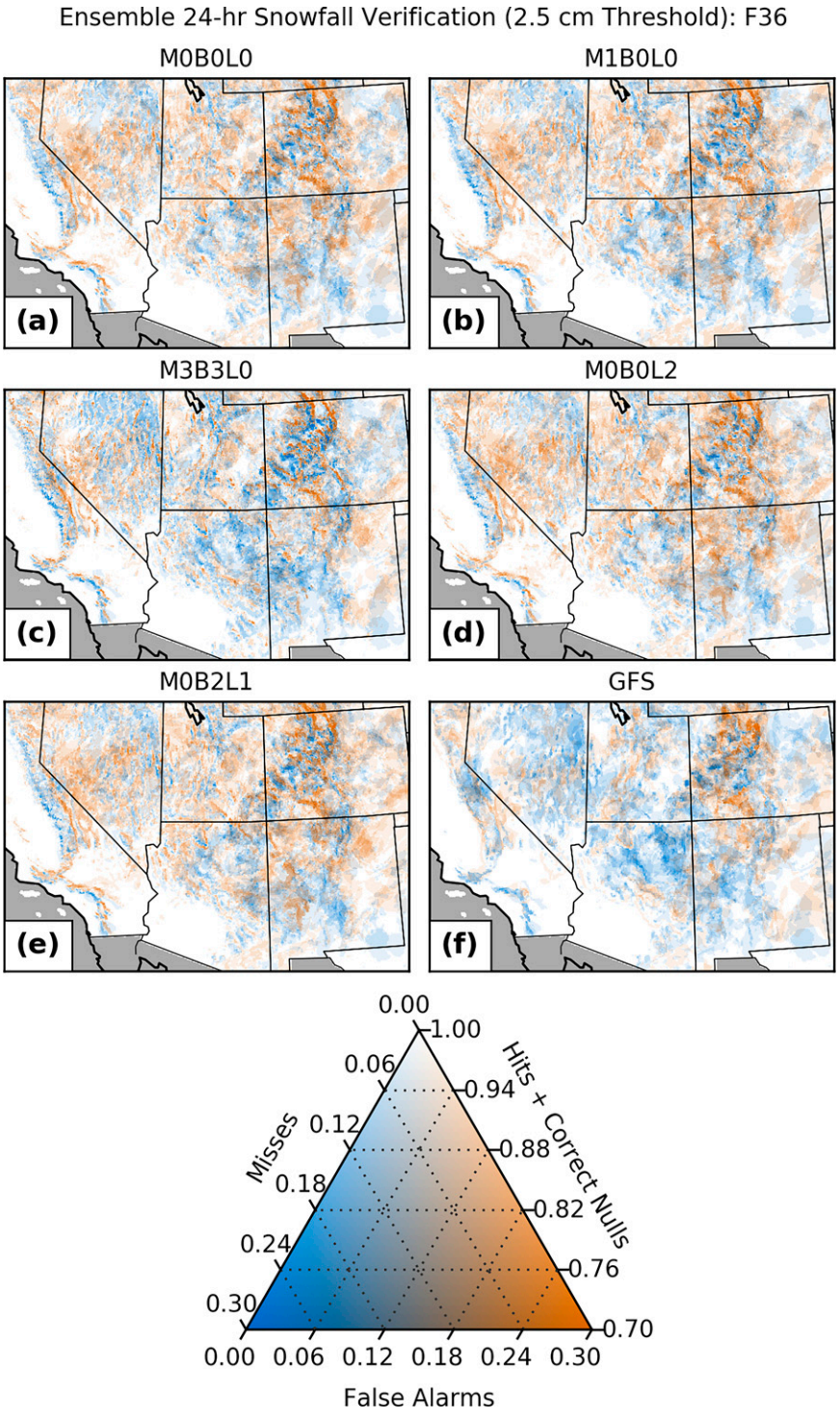

FIG. 14. As in Fig. 8, but for 24-h snowfall with a threshold of 2.5 cm.

In comparing the GFS to the CAMs, the surface sensible weather parameters are generally better (a lower bias and lower RMSE) in the GFS, while the precipitation fields are generally better (lower bias with comparable ETS) in the CAMs. This suggests that the CAMs have an inherent advantage over lower-resolution models with respect to predicting precipitation, even in the cool season, despite the lower importance of convection compared to the warm season. In addition, we speculate that for surface fields, particularly outside of mountainous areas, the forecasts are more sensitive to the physics choices than to the grid spacing.

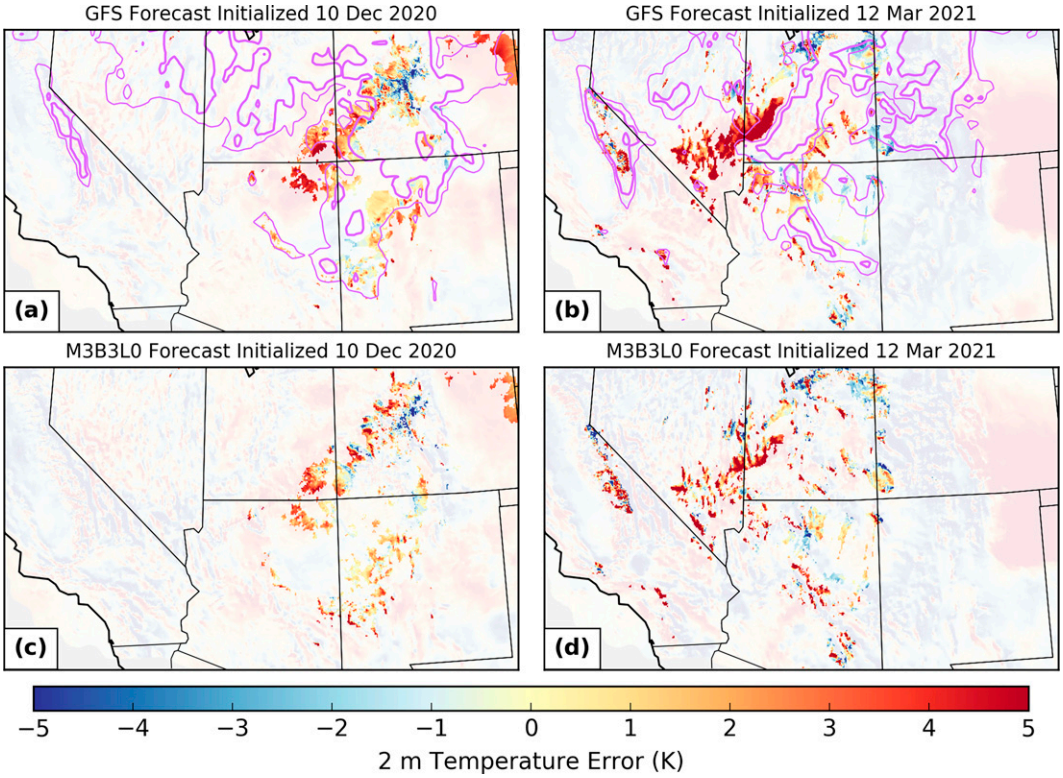

FIG. 15. The 24-h forecast 2-m temperature errors (in color fills) from the (a),(c) 10 Dec and (b),(d) 12 Mar cases. Forecasts from the (top) the GFS and (bottom) M3B3L0 are shown. The regions where each forecast misses day 1 (forecast hours 12–36) 24-h snowfall accumulations of 2.5 cm or more are highlighted, and all other regions are covered with a semitransparent mask. In addition, the forecast 0° and 3°C 2-m temperature contours are given by the thick and thin purple lines. The 2-m temperature contours are omitted from (c) and (d) for legibility.

As mentioned above, a goal of this work is to provide recommendations for physics parameterizations the future RRFS ensemble. Overall, the Thompson and NSSL microphysics perform well for precipitation and snowfall prediction in the cool season. The lower bias in precipitation the NSSL scheme is the result of two offsetting biases: the domain wide over prediction of precipitation and under prediction in a specific region. Therefore, it is not clear that one scheme is universally better than the other, and both are recommended for use. Furthermore, according to these results, the MYNN PBL scheme performs the best overall out of all the PBL schemes tested, and the Shin–Hong PBL scheme is generally comparable to MYNN in terms of precipitation and surface fields; both are recommended for use. The TKE-EDMF PBL scheme's overprediction of precipitation suggests the need for improvement. This is important given that TKE-EDMF has replaced K-EDMF as the PBL scheme used in the operational GFSv16. Finally, the NoahMP LSM results in lower precipitation biases in higher ETS in the forecasts, but the Noah LSM performs the best for the surface fields. Therefore, both are recommended for use in the future RRFS ensemble, depending on which field is deemed more important to be unbiased.

Also, the evaluation herein focuses on the performance of individual forecasts. Future work will emphasize performance of the overall ensemble, when initial and boundary condition perturbations are also included. Ensemble consensus products, such as the mean and localized probability matched mean (Snook et al. 2019; Clark 2017), as well as ensemble evaluation metrics, such as rank histograms (Hamill 2001; see e.g., Duda et al. 2014), attributes diagrams, and area under the relative operating characteristic (ROC) curve (see e.g., Loken et al. 2017) will be evaluated to assess the ensemble forecasting performance, given the goal toward optimal design of the operational RRFS ensemble. The forecasts produced during the HMT FFaIR experiment by this group for warm-season precipitation forecasting will be evaluated in similar ways.

Acknowledgments. This work was primarily supported by the NOAA UFS R2O program under Grant NA16OAR4320115 and the NOAA Testbed Program OAR Grant NA19OAR4590141 to CAPS. All forecasts were run on the Texas Advanced Computing Center (TACC) Frontera supercomputer, with allocations obtained through the NSF XSEDE program. Additional thanks to Benjamin Blake at NOAA EMC for sharing their FV3-LAM configurations and static files. Thanks also to James Nelson, Kirstin Harnos, James Correia, and Benjamin Albright of NOAA/WPC for facilitating the HMT Winter Weather Experiments and providing feedback on the forecast configurations. Finally, thanks to

three anonymous reviewers whose comments increased the clarity of this paper.

Data availability statement. The FV3-LAM model output evaluated herein is archived at the Center for Analysis and Prediction of Storms (CAPS) and is available for download at https://doi.org/10.15763/DBS.CAPS.001, and GFS forecasts are available from UCAR's Research Data Archive (RDA) repository (https://rda.ucar.edu/datasets/ds084.1/). Stage-IV precipitation analyses can be downloaded from UCAR's Earth Observing Laboratory archive (https://data.eol.ucar.edu/dataset/21.093), National Operational Hydrologic Remote Sensing Center (NOHRSC) snowfall analyses are available from NOHRSC (https://www.nohrsc.noaa.gov/snowfall_v2/data/), and GFS final analyses are available from the UCAR RDA repository (https://rda.ucar.edu/datasets/ds083.3/). Unrestricted Mesoscale Analysis (URMA) data were provided by NOAA EMC and are available for download at https://doi.org/10.15763/DBS.CAPS.001.

REFERENCES

Aligo, E. A., B. Ferrier, and J. R. Carley, 2018: Modified NAM microphysics for forecasts of deep convective storms. *Mon. Wea. Rev.*, **146**, 4115–4153, https://doi.org/10.1175/MWR-D-17-0277.1.

Baldwin, M. E., and J. S. Kain, 2006: Sensitivity of several performance measures to displacement error, bias, and event frequency. *Wea. Forecasting*, **21**, 636–648, https://doi.org/10.1175/WAF933.1.

Barthold, F. E., T. E. Workoff, B. A. Cosgrove, J. J. Gourley, D. R. Novak, and K. M. Mahoney, 2015: Improving flash flood forecasts: The HMT-WPC flash flood and intense rainfall experiment. *Bull. Amer. Meteor. Soc.*, **96**, 1859–1866, https://doi.org/10.1175/BAMS-D-14-00201.1.

Benjamin, S. G., J. M. Brown, and T. G. Smirnova, 2016: Explicit precipitation-type diagnosis from a model using a mixed-phase bulk cloud-precipitation microphysics parameterization. *Wea. Forecasting*, **31**, 609–619, https://doi.org/10.1175/WAF-D-15-0136.1.

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

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

Black, T. L., and Coauthors, 2021: A limited area modeling capability for the finite-volume cubed-sphere (FV3) dynamical core and comparison with a global two-way nest. *J. Adv. Model. Earth Syst.*, **13**, e2021MS002483, https://doi.org/10.1029/2021MS002483.

Bytheway, J. L., C. D. Kummerow, and C. Alexander, 2017: A features-based assessment of the evolution of warm season precipitation forecasts from the HRRR model over three years of development. *Wea. Forecasting*, **32**, 1841–1856, https://doi.org/10.1175/WAF-D-17-0050.1.

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

Clark, A. J., 2017: Generation of ensemble mean precipitation forecasts from convection-allowing ensembles. *Wea. Forecasting*, **32**, 1569–1583, https://doi.org/10.1175/WAF-D-16-0199.1.

——, W. A. Gallus, and M. L. Weisman, 2010: Neighborhood-based verification of precipitation forecasts from convection-allowing NCAR WRF Model simulations and the operational NAM. *Wea. Forecasting*, **25**, 1495–1509, https://doi.org/10.1175/2010WAF2222404.1.

——, and Coauthors, 2020: A real-time, simulated forecasting experiment for advancing the prediction of hazardous convective weather. *Bull. Amer. Meteor. Soc.*, **101**, E2022–E2024, https://doi.org/10.1175/BAMS-D-19-0298.1.

Dawson, D. T., M. Xue, J. A. Milbrandt, and M. K. Yau, 2010: Comparison of evaporation and cold pool development between single-moment and multimoment bulk microphysics schemes in idealized simulations of tornadic thunderstorms. *Mon. Wea. Rev.*, **138**, 1152–1171, https://doi.org/10.1175/2009MWR2956.1.

De Pondeca, M. S. F. V., and Coauthors, 2011: The real-time mesoscale analysis at NOAA's National Centers for Environmental Prediction: Current status and development. *Wea. Forecasting*, **26**, 593–612, https://doi.org/10.1175/WAF-D-10-05037.1.

Duda, J. D., and D. D. Turner, 2021: Large-sample application of radar reflectivity object-based verification to evaluate HRRR warm-season forecasts. *Wea. Forecasting*, **36**, 805–821, https://doi.org/10.1175/WAF-D-20-0203.1.

——, X. Wang, F. Kong, and M. Xue 2014: Using varied microphysics to account for uncertainty in warm-season QPF in a convection-allowing ensemble. *Mon. Wea. Rev.*, **142**, 2198–2219, https://doi.org/10.1175/MWR-D-13-00297.1.

Ek, M. B., K. E. Mitchell, Y. Lin, E. Rogers, P. Grunmann, V. Koren, G. Gayno, and J. D. Tarpley, 2003: Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. *J. Geophys. Res.*, **108**, 8851, https://doi.org/10.1029/2002JD003296.

Fairall, C. W., E. F. Bradley, J. S. Godfrey, G. A. Wick, J. B. Edson, and G. S. Young, 1996: Cool-skin and warm-layer effects on sea surface temperature. *J. Geophys. Res.*, **101**, 1295–1308, https://doi.org/10.1029/95JC03190.

Firl, G., L. Carson, M. Harrold, L. Bernardet, and D. Heinzeller, 2021: Common Community Physics Package Single Column Model (SCM). Developmental Testbed Center, 44 pp., https://dtcenter.org/GMTB/v5.0.0/scm-ccpp-guide-v5.0.0.pdf.

Flora, M. L., P. S. Skinner, C. K. Potvin, A. E. Reinhart, T. A. Jones, N. Yussouf, and K. H. Knopfmeier, 2019: Object-based verification of short-term, storm-scale probabilistic mesocyclone guidance from an experimental Warn-on-Forecast system. *Wea. Forecasting*, **34**, 1721–1739, https://doi.org/10.1175/WAF-D-19-0094.1.

Gallo, B. T., and Coauthors, 2021: Exploring convection-allowing model evaluation strategies for severe local storms using the finite-volume cubed-sphere (FV3) model core. *Wea. Forecasting*, **36**, 3–19, https://doi.org/10.1175/WAF-D-20-0090.1.

Ganai, M., S. Tirkey, R. P. M. Krishna, and P. Mukhopadhyay, 2021: The impact of modified rate of precipitation conversion parameter in the convective parameterization scheme of operational weather forecast model (GFS T1534) over Indian

summer monsoon region. *Atmos. Res.*, **248**, 105185, https://doi.org/10.1016/j.atmosres.2020.105185.

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

——, ——, ——, ——, ——, ——, T. L. Jensen, and J. K. Wolff, 2017b: Seasonal analysis of cloud objects in the High-Resolution Rapid Refresh (HRRR) model using object-based verification. *J. Appl. Meteor. Climatol.*, **56**, 2317–2334, https://doi.org/10.1175/JAMC-D-17-0004.1.

——, ——, S. E. Nebuda, T. L. Jensen, P. S. Skinner, E. Gilleland, T. A. Supinie, and M. Xue, 2021: Evaluating the impact of planetary boundary layer, land surface model, and microphysics parameterization schemes on cold cloud objects in simulated *GOES-16* brightness temperatures. *J. Geophys. Res.*, **126**, e2021JD034709, https://doi.org/10.1029/2021JD034709.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167, https://doi.org/10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2.

——, 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560, https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.

Han, J., and C. S. Bretherton, 2019: TKE-based moist eddy-diffusivity mass-flux (EDMF) parameterization for vertical turbulent mixing. *Wea. Forecasting*, **34**, 869–886, https://doi.org/10.1175/WAF-D-18-0146.1.

——, M. L. Witek, J. Teixeira, R. Sun, H.-L. Pan, J. K. Fletcher, and C. S. Bretherton, 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. *Wea. Forecasting*, **31**, 341–352, https://doi.org/10.1175/WAF-D-15-0053.1.

Harnos, K., J. Correia, B. Albright, M. Bodner, and J. Nelson, 2021: 11th Annual Winter Weather Experiment: Findings and results. NOAA, 50 pp., https://origin.wpc.ncep.noaa.gov/hmt/wwe2021/11th_Annual_HMT_WWE_Final_Report.pdf.

Harris, L., L. Zhou, X. Chen, and J.-H. Chen, 2020: The nonhydrostatic solver of the GFDL finite-volume cubed-sphere dynamical core. NOAA Tech. Memo. OAR GFDL, 2020-003, 7 pp., https://repository.library.noaa.gov/view/noaa/27489.

Ikeda, K., M. Steiner, J. Pinto, and C. Alexander, 2013: Evaluation of cold-season precipitation forecasts generated by the hourly updating high-resolution Rapid Refresh model. *Wea. Forecasting*, **28**, 921–939, https://doi.org/10.1175/WAF-D-12-00085.1.

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

Jiang, M., and Coauthors, 2017: Potential influences of neglecting aerosol effects on the NCEP GFS precipitation forecast. *Atmos. Chem. Phys.*, **17**, 13 967–13 982, https://doi.org/10.5194/acp-17-13967-2017.

Lin, S. J., 2004: A "vertically Lagrangian" finite-volume dynamical core for global models. *Mon. Wea. Rev.*, **132**, 2293–2307, https://doi.org/10.1175/1520-0493(2004)132<2293:AVLFDC>2.0.CO;2.

Loken, E. D., A. J. Clark, M. Xue, and F. Kong, 2017: Comparison of next-day probabilistic severe weather forecasts from coarse- and fine-resolution CAMs and a convection-allowing ensemble. *Wea. Forecasting*, **32**, 1403–1421, https://doi.org/10.1175/WAF-D-16-0200.1.

Long, P. E., 1986: An economic and compatible scheme for parameterizing the stable surface layer in the medium range forecast model. NCEP Office Note 321, 24 pp.

——, 1990: Derivation and suggested method of the application of simplified relations for surface fluxes in the medium-range forecast model: Unstable case. NCEP Office Note 356, 53 pp.

Manikin, G. S., 2005: An overview of precipitation type forecasting using NAM and SREF data. Preprints, *21st Conf. on Weather Analysis and Forecasting/17th Conf. on Numerical Weather Prediction*, Washington, DC, Amer. Meteor. Soc., 8A.6 https://ams.confex.com/ams/pdfpapers/94838.pdf.

Mansell, E. R., and C. L. Ziegler, 2013: Aerosol effects on simulated storm electrification and precipitation in a two-moment bulk microphysics model. *J. Atmos. Sci.*, **70**, 2032–2050, https://doi.org/10.1175/JAS-D-12-0264.1.

——, ——, and E. C. Bruning, 2010: Simulated electrification of a small thunderstorm with two-moment bulk microphysics. *J. Atmos. Sci.*, **67**, 171–194, https://doi.org/10.1175/2009JAS2965.1.

Mason, I. B., 2003: Binary events. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, I. T. Jolliffe and D. B. Stephenson, Eds., Wiley, 37–76.

McMillen, J. D., and W. J. Steenburgh, 2015: Capabilities and limitations of convection-permitting WRF simulations of lake-effect systems over the Great Salt Lake. *Wea. Forecasting*, **30**, 1711–1731, https://doi.org/10.1175/WAF-D-15-0017.1.

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

Morrison, H., and A. Gettelman, 2008: A new two-moment bulk stratiform cloud microphysics scheme in the Community Atmosphere Model, version 3 (CAM3). Part I: Description and numerical tests. *J. Climate*, **21**, 3642–3659, https://doi.org/10.1175/2008JCLI2105.1.

Nakanishi, M., and H. Niino, 2009: Development of an improved turbulence closure model for the atmospheric boundary layer. *J. Meteor. Soc. Japan*, **87**, 895–912, https://doi.org/10.2151/jmsj.87.895.

Nelson, B. R., O. P. Prat, D. J. Seo, and E. Habib, 2016: Assessment and implications of NCEP stage IV quantitative precipitation estimates for product intercomparisons. *Wea. Forecasting*, **31**, 371–394, https://doi.org/10.1175/WAF-D-14-00112.1.

Niu, G.-Y., and Coauthors, 2011: The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *J. Geophys. Res.*, **116**, D12109, https://doi.org/10.1029/2010JD015139.

Olson, J. B., J. S. Kenyon, W. M. Angevine, J. M. Brown, M. Pagowski, and K. Sušelj, 2019: A description of the MYNN-EDMF scheme and the coupling to other components in WRF-ARW. NOAA Tech. Memo. OAR GSD-61, 42 pp., https://doi.org/10.25923/n9wm-be49.

——, T. Smirnova, J. S. Kenyon, D. D. Turner, J. M. Brown, W. Zheng, and B. W. Green, 2021: A description of the MYNN surface-layer scheme. NOAA Tech. Memo. OAR GSL-67, 26 pp., https://doi.org/10.25923/f6a8-bc75.

Pondeca, M., S. Levine, J. Carley, Y. Lin, Y. Zhu, and J. Purser, 2015: Ongoing improvements to the NCEP Real Time Mesoscale Analysis (RTMA) and UnRestricted Mesoscale

Analysis (URMA) and NCEP/EMC. NOAA, 2 pp., https://www.wcrp-climate.org/WGNE/BlueBook/2015/individual-articles/01_Pondeca_Manuel_etal_RTMA.pdf.

Potvin, C. K., and Coauthors, 2019: Systematic comparison of convection-allowing models during the 2017 NOAA HWT Spring Forecasting Experiment. *Wea. Forecasting*, **34**, 1395–1416, https://doi.org/10.1175/WAF-D-19-0056.1.

Putman, W. M., and S. J. Lin, 2007: Finite-volume transport on various cubed-sphere grids. *J. Comput. Phys.*, **227**, 55–78, https://doi.org/10.1016/j.jcp.2007.07.022.

Roberts, B., B. T. Gallo, I. L. Jirak, A. J. Clark, D. C. Dowell, X. Wang, and Y. Wang, 2020: What does a convection-allowing ensemble of opportunity buy us in forecasting thunderstorms? *Wea. Forecasting*, **35**, 2293–2316, https://doi.org/10.1175/WAF-D-20-0069.1.

Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, https://doi.org/10.1175/2007MWR2123.1.

Schwartz, C. S., and Coauthors, 2009: Next-day convection-allowing WRF Model guidance: A second look at 2-km versus 4-km grid spacing. *Mon. Wea. Rev.*, **137**, 3351–3372, https://doi.org/10.1175/2009MWR2924.1.

Shin, H. H., and S. Y. Hong, 2015: Representation of the subgrid-scale turbulent transport in convective boundary layers at gray-zone resolutions. *Mon. Wea. Rev.*, **143**, 250–271, https://doi.org/10.1175/MWR-D-14-00116.1.

Skinner, P. S., and Coauthors, 2018: Object-based verification of a prototype Warn-on-Forecast System. *Wea. Forecasting*, **33**, 1225–1250, https://doi.org/10.1175/WAF-D-18-0020.1.

Smirnova, T. G., J. M. Brown, S. G. Benjamin, and J. S. Kenyon, 2016: Modifications to the Rapid Update Cycle Land Surface Model (RUC LSM) available in the Weather Research and Forecasting (WRF) Model. *Mon. Wea. Rev.*, **144**, 1851–1865, https://doi.org/10.1175/MWR-D-15-0198.1.

Snook, N., F. Kong, K. A. Brewster, M. Xue, K. W. Thomas, T. A. Supinie, S. Perfater, and B. Albright, 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. *Wea. Forecasting*, **34**, 781–804, https://doi.org/10.1175/WAF-D-18-0155.1.

Sobash, R. A., and J. S. Kain, 2017: Seasonal variations in severe weather forecast skill in an experimental convection-allowing model. *Wea. Forecasting*, **32**, 1885–1902, https://doi.org/10.1175/WAF-D-17-0043.1.

——, G. S. Romine, C. S. Schwartz, D. J. Gagne, and M. L. Weisman, 2016: Explicit forecasts of low-level rotation from convection-allowing models for next-day tornado prediction. *Wea. Forecasting*, **31**, 1591–1614, https://doi.org/10.1175/WAF-D-16-0073.1.

Tessendorf, S. A., and Coauthors, 2021: Differentiating freezing drizzle and freezing rain in HRRR model forecasts. *Wea. Forecasting*, **36**, 1237–1251, https://doi.org/10.1175/WAF-D-20-0138.1.

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

——, P. R. Field, R. M. Rasmussen, and W. D. Hall, 2008: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization. *Mon. Wea. Rev.*, **136**, 5095–5115, https://doi.org/10.1175/2008MWR2387.1.

Zhang, C., and Coauthors, 2019: How well does an FV3-based model predict precipitation at a convection-allowing resolution? Results from CAPS forecasts for the 2018 NOAA Hazardous Weather Test Bed with different physics combinations. *Geophys. Res. Lett.*, **46**, 3523–3531, https://doi.org/10.1029/2018GL081702.

Zhu, K., and Coauthors, 2018: Evaluation of real-time convection-permitting precipitation forecasts in China during the 2013–2014 summer season. *J. Geophys. Res. Atmos.*, **123**, 1037–1064, https://doi.org/10.1002/2017JD027445.