Towards a Unified Setup to Simulate Mid-Latitude and Tropical Mesoscale Convective Systems at Kilometer-Scales

Andreas F. Prein1 Ⓢ, Ming Ge1, Alexandra Ramos Valle1, Dié Wang2, and Scott E. Giangrande2 Ⓢ

1National Center for Atmospheric Research, Boulder, CO, USA, 2Brookhaven National Laboratory, Upton, NY, USA

Abstract Mesoscale convective systems (MCSs) are the main source of precipitation in the tropics and parts of the mid-latitudes and are responsible for high-impact weather worldwide. Studies showed that deficiencies in simulating mid-latitude MCSs in state-of-the-art climate models can be alleviated by kilometer-scale models. However, whether these models can also improve tropical MCSs and whether we can find model settings that perform well in both regions is understudied. We take advantage of high-quality MCS observations collected over the Atmospheric Radiation Measurement (ARM) facilities in the US Southern Great Plains (SGP) and the Amazon basin near Manaus (MAO) to evaluate a perturbed physics ensemble of simulated MCSs with 4 km horizontal grid spacing. A new model evaluation method is developed that enables to distinguish biases stemming from spatiotemporal displacements of MCSs from biases in their reflectivity and cloud shield. Amazon MCSs are similarly well simulated across these evaluation metrics than SGP MCSs despite the challenges anticipated from weaker large-scale forcing in the tropics. Generally, SGP MCSs are more sensitive to the choice of model microphysics, while Amazon cases are more sensitive to the planetary boundary layer (PBL) scheme. Although our tested model physics combinations had strengths and weaknesses, combinations that performed well for SGP simulations result in worse results in the Amazon basin and vice versa. However, we identified model settings that perform well at both locations, which include the Thompson and Morrison microphysics coupled with the Yonsei University (YSU) PBL scheme and the Thompson scheme coupled with the Mellor-Yamada-Janic PBL scheme.

1. Introduction

Mesoscale convective systems (MCSs) play an important role in Earth’s energy balance (Houze, 2018) and are essential for Earth’s water cycle in the tropics (Feng et al., 2016) and mid-latitude regions (Feng et al., 2016; Feng, Leung, et al., 2021; Fritsch et al., 1986). These systems are prolific rain producers and are the main cause of warm-season flooding (Pokharel et al., 2018; Rasmussen et al., 2014; R. S. Schumacher & Johnson, 2005). Observations of MCSs over the continental US indicate that extreme precipitation rates associated with MCSs have significantly increased during the past decades (Feng et al., 2016) and MCSs are predicted to further intensify in the future climate (Prein et al., 2017). Nevertheless, a major bottleneck for predicting possible climate change effects on climate extremes has been related to the poor representation of MCS intensity (e.g., precipitation rates, updraft strength) and spatiotemporal evolution in state-of-the-art models (Donner et al., 2016; Lin et al., 2022; Wang et al., 2020). Improving our MCS modeling capabilities on weather, seasonal, and climate time scales is essential to advance the credibility of model predictions.

The frontier of global and regional atmospheric modeling has reached convection-permitting scales (horizontal grid spacings Δx ≤ 4 km; Satoh et al., 2019). While global large eddy (LES) simulations on climate timescales are far out of reach, convection-permitting “gray-zone” decadal simulations are already feasible in regional models (e.g., Berthou et al., 2020; Liu et al., 2017) and will soon be achievable with global models (e.g., Stevens et al., 2019). Convection-permitting models (CPMs) can explicitly simulate deep convective clouds, which revolutionizes our ability to simulate and predict severe weather and climate extremes (Clark et al., 2016; Prein et al., 2015). These CPMs substantially improve the simulation of MCSs including their propagation direction and speed, evolution, size, and associated extreme precipitation (Prein et al., 2020). Although this progress is encouraging, CPMs have difficulties simulating MCS cold pool and draft dynamics that modulate the MCS lifecycle and development, especially in weakly forced environments (Haberlie & Ashley, 2019a; Wang et al., 2020).

In this study, we investigate the ability of the Advanced Research Weather Research and Forecasting (AR-WRF or short WRF) model (Powers et al., 2017; Skamarock & Klemp, 2008) at convection-permitting resolution (4 km
horizontal grid spacing) to simulate MCSs that overpassed the US Department of Energy's (DOE) Atmospheric Radiation Measurement (ARM) (Mather & Voyles, 2013) sites in the US Southern Great Plains (SGP, Lamont, Oklahoma) (Sisterson et al., 2016) and the Amazon basin (MAO, Manaus, Brazil) (Martin et al., 2016). MCSs in those regions initiate and develop under very different environmental conditions, which promotes distinct convection lifecycle characteristics and associated precipitation behaviors (Wang et al., 2019).

Moreover, Wang et al. (2019) recently showed that MCSs over the SGP and MAO sites feature similar rainfall rates and accumulations, though having larger stratiform rainfall contributions in the US SGP events. Similarly, the convective cold pools have comparable strength in both regions. However SGP MCSs show more intense convective updrafts, deeper and stronger convective downdrafts, and larger mass flux than MAO MCSs (Wang et al., 2020). The SGP MCS investigated for this study predominantly occur in the springtime (see Table 1) and feature strong large-scale forcing (e.g., frontal passages), which is distinctly different from MAO MCSs that typically occur under much weaker synoptic-scale forcing. These differences in synoptic-scale and thermodynamic forcings motivate this study to explore the skill of CPMs and their sensitivities to model physics in simulating MCSs in both regions. There have been several studies that investigate the sensitivity of simulated MCSs to the model microphysics (Fan et al., 2017; Feng et al., 2018; Han et al., 2019; Shpund et al., 2019; Xue et al., 2017) and planetary boundary layer (PBL) schemes (Stephan & Alexander, 2014) using kilometer-scale models. Most of these studies focus on US mid-latitude MCSs and do not assess sensitivities across climate zones.

The MCS overpasses over the SGP and MAO sites that are described in Wang et al. (2019) provide a unique opportunity for in-depth analyses of the performance of CPMs in simulating MCSs in these two regimes. However, using ARM observations for model evaluation is challenging due to the limited spatial coverage of the observations and spatiotemporal displacements of MCSs in the simulations. Traditional evaluation methods cannot be used in such situations due to the so-called “double-penalty problem” (Prein et al., 2013; Roberts & Lean, 2008). In these cases, simulation performance is penalized twice compared to null behaviors, once for not properly simulating the primary event and once for including secondary features that were not observed. We present a new method that identifies the spatiotemporal displacement of simulated storms, which allows us to disentangle displacement errors from other modeling errors. We use this method to test the sensitivity of simulated MCSs at the SGP and MAO sites to perturbed model PBL- and micro-physics schemes.

2. Data and Methods

2.1. MCS Case Selection

MCSs in the US Great Plains are well studied using observations (Fritsch et al., 1986; Haberlie & Ashley, 2019b; Houze, 2018; Song et al., 2019; Wang et al., 2019) and models (Feng et al., 2018; Prein et al., 2020; Trier et al., 2010). US MCSs have a strong seasonality and are most frequent in the southern Great Plains during spring and in the northern Great Plains during summer (Li et al., 2021). Spring MCSs are frequently related to frontal passages, feature an enhanced Great Plains low-level jet, and are typically squall lines (Song et al., 2019). Spring MCSs occur under large convective available potential energy (CAPE) and convective inhibition (CIN) anomalies, while summertime MCSs occur under much weaker synoptic-scale forcing (Feng, Song, et al., 2021). Even kilometer-scale models can have difficulties in simulating MCSs under weakly forced summertime conditions (Prein et al., 2020).

MCSs are the dominant rain-producer in the Amazon basin, which features about 7,200 MCSs per year (Anselmo et al., 2021; Rehbein et al., 2018) contributing approximately 50% of the annual rainfall (Feng, Leung, et al., 2021). MCSs in the Amazon can have various morphologies and sizes but typically develop in lines (Anselmo et al., 2021). The number of MCSs during the ARM Green Ocean Amazon (GoAmazon) campaign period (2014–2015), whose data we leverage in this study, was about 50% lower than the 2000–2013 climatology (Rehbein et al., 2019) probably due to positive equatorial Pacific sea surface temperatures and reduced moisture transport into the Amazon basin. For additional context, the larger-scale regimes associated with GoAmazon mature MCS events are summarized by Wang et al. (2019) and more generically in Giangrande et al. (2020).

We investigate MCS events over the US DOE ARM (Mather & Voyles, 2013) SGP (Sisterson et al., 2016) site and the 2014/15 GoAmazon field campaign site near Manaus Brazil (MAO; Giangrande et al., 2017; Martin et al., 2016, 2017). In previous work, we identified 16 MCS overpasses over the SGP site and 44 over the MAO sites (Wang et al., 2019). From these cases, we simulate a subset of 13 cases at the SGP site and 41 cases at
Table 1
Selected MCS Cases in the US Central Great Plains (SGP) and Amazon Basin (MAO)

| Region | Date and time [UTC] | Season | Morphology               |
|--------|---------------------|--------|--------------------------|
| SGP    | 2012.05.31 04:00    | Spring | Squall line              |
| SGP    | 2012.06.15 07:00    | Spring | Squall line              |
| SGP    | 2013.05.09 07:00    | Spring | Squall line              |
| SGP    | 2013.06.05 09:00    | Spring | Bow echo                 |
| SGP    | 2013.06.17 07:00    | Spring | Squall line              |
| SGP    | 2014.06.02 04:00    | Spring | Squall line              |
| SGP    | 2014.06.05 12:00    | Spring | Bow echo                 |
| SGP    | 2014.06.12 06:00    | Spring | Squall line              |
| SGP    | 2014.06.28 16:00    | Spring | Weakly organized         |
| SGP    | 2014.07.10 10:00    | Summer | Training line            |
| SGP    | 2016.03.08 15:00    | Spring | Weak squall line         |
| SGP    | 2016.06.18 10:00    | Spring | Mesoscale convective complex |
| SGP    | 2016.07.29 09:00    | Summer | Squall line              |
| MAO    | 2014.08.16 14:00    | Dry    | Local, small system      |
| MAO    | 2014.09.17 17:00    | Dry    | Squall line              |
| MAO    | 2015.06.21 14:00    | Dry    | Squall line              |
| MAO    | 2014.04.01 15:00    | Wet    | Training line            |
| MAO    | 2014.12.10 14:00    | Wet    | Local, weakly organized  |
| MAO    | 2015.03.28 15:00    | Wet    | Local, weakly organized  |
| MAO    | 2015.04.12 12:00    | Wet    | Squall line              |
| MAO    | 2014.10.04 13:00    | Transition | Squall line              |
| MAO    | 2014.10.18 14:00    | Transition | Local, weakly organized |
| MAO    | 2014.11.17 18:00    | Transition | Squall line              |
| MAO    | 2015.11.06 12:00    | Transition | Squall line              |

Note: The date and time indicates the overpass of the MCS over the corresponding ARM site as defined in Wang et al. (2019).

2.2. Observational Data Sets

We use S-band Z (10-cm wavelength, 3 GHz) and satellite-derived cloud BT for model evaluation. For MAO MCS cases, volumetric radar observations collected during the GoAmazon2014/15 campaign by the SIPAM S-band radars are used. These data are quality controlled and interpolated to a three-dimensional grid (C. Schumacher & Funk, 2018). Data are available every 12 min on a 2 km horizontal grid, which covers a maximum area of 480 × 480 km. The vertical grid spacing is 500 m at constant altitude plan position indicator (CAPPI) with a vertical grid spacing of 500 m up to 20 km height. We test the sensitivity of our model evaluation using scans at 2, 4, and 6 km altitudes. Those levels are chosen due to the complete data coverage over the radar site (higher levels have missing data over the radar location). The 2 km CAPPIs (those that sample to the furthest distance) are limited to 180 km in range, which informed our decision about the default scanning distance that we will introduce in Section 2.4. For the SGP MCS events, we use the GridRad (Homeyer & Bowman, 2017) product that merges Z from the US National Weather Service NEXRAD WSR-88D high-resolution S-band Doppler weather radars (Ansari et al., 2018). GridRad provides hourly, three-dimensional, Z on an 0.02° horizontal and 1 km vertical grid between 1995 and 2018 covering the Contiguous United States. We use the same scanning distance as for the Amazon MCS evaluation for consistency, but the larger radar coverage in the US also allows testing the sensitivity of our analysis to the scanning distance setting. The NEXRAD and SIPAM are quality controlled and accurate within 1–2 dBz (Wang et al., 2019).

The BT observations are derived from the GOES-13 satellite Channel 4 (Knapp & Wilkins, 2018). Channel 4 measurements (often referred to as the “Cirrus” Band; approx. central wavelength is 1.37 μm) were selected owing to their high contrast in identifying anvil clouds and their good agreement with simulated BT data (Feng, Leung, et al., 2021). The only exception where we substituted GOES13 data with GOES15 data is 5 June 2013, MCS case that over-passed the SGP site due to missing data in GOES13. GOES data are provided hourly with all data mapping to the nearest hour on an equal angle grid with a spacing of 0.04°.

2.3. Model Setup

We use the WRF model version 4.1.5 (Powers et al., 2017; Skamarock & Klemp, 2008) for the simulations. The simulation domains are shown in Figure 1. Each domain consists of 500 × 500 grid cells and 96 stretched vertical levels. The horizontal grid spacing is approximately 4 km, which is sufficient to capture bulk MCS properties reasonably well at computational affordable costs (Prein et al., 2020). Each simulation is started 24-hr before an
MCS overpass at the corresponding ARM site (Wang et al., 2019) and has a total runtime of 36-hr. The initial and lateral boundary conditions are derived from hourly pressure level data from the fifth generation European Center for Medium-Range Weather Forecasting (ECMWF) reanalysis (ERA5; Hersbach et al., 2020). We use the Noah-MP land surface model (Niu et al., 2011), the RRTMG shortwave and longwave radiation scheme (Iacono et al., 2008) and do not use a convection parameterization scheme. We test three options each for the microphysics and PBL parameterizations, to be described below. These schemes were selected as they represent several of the most widely tested/used options and feature different levels of complexity and underlying assumptions.

For our microphysical sensitivity tests, we chose the Thompson (Thompson et al., 2004), the Morrison (Morrison et al., 2009), and the P3 (Morrison & Milbrandt, 2015) schemes. All of these are bulk microphysics schemes that vary in their representation of hydrometeors. The Thompson scheme uses two moments for cloud ice and rain hydrometeors and one moment for graupel/hail, rain, and snow; which allows it to predict ice and rain density. However, Thompson representations for snow, rain, and graupel/hail properties is otherwise limited compared to the other schemes tested. The Morrison microphysics scheme is more complex than the Thompson scheme since it also represents two moments of snow and graupel/hail. The P3 scheme follows the full 2-moment implementation of the Morrison scheme but includes a detailed prediction of ice particle properties (conceptually similar to Jensen et al. [2017]). This change avoids the artificial classification of frozen hydrometeors into ice, snow, and graupel/hail categories. This scheme has a conceptual advantage over the Morrison and Thompson schemes but is less widely used and tested.

For the PBL parameterization sensitivity testing, we considered the Yonsei University (YSU; Hong et al., 2006), the Mellor-Yamada-Janjić (MYJ; Janjić, 1994; Mesinger, 1993), and Mellor-Yamada Nakanishi Niino Level 2.5 schemes (MYNN2.5; Nakanishi & Niino, 2006, 2009). YSU is a non-local scheme that uses a first-order closure and has an improved simulation of deeper vertical mixing in buoyancy-driven PBLs and shallower mixing in strong-wind regimes compared to successor PBL schemes (e.g., the MRF scheme; Hong & Pan, 1996). However, YSU features systematic biases that may include issues such as PBLs that deepen too vigorously for springtime deep convective environments, resulting in an underestimation of near-surface buoyancy (Coniglio et al., 2013). In contrast, the MYJ parameterization is a local 1.5-order closure scheme, which improves PBL simulations compared to its preceding schemes (Mellor & Yamada, 1982), without increases in computational costs. However, MYJ tends to undermix PBLs for locations upstream of convection (Coniglio et al., 2013). Finally, we employ

Figure 1. Computational 4 km WRF domains for simulating MCS cases in the US Southern Great Plains (SGP; a) and Amazon basin (MAO; b). The colored contours show the model topography. Each domain consists of 300 × 300 × 96 grid cells. The black circle shows the location of the ARM SGP and MAO site.
the MYNN2.5, which is a local scheme that uses a 1.5-order closure and improves the PBL depiction compared to non-local schemes (e.g., YSU) during springtime in environments that support deep convection (Coniglio et al., 2013). Similar to MYJ, the local formulations of MYNN2 do not fully account for deep vertical mixing.

BT needs to be estimated from WRF output since it is not a standard output variable, either by running a radiative transfer model (NCEP, 2020) or by applying empirical relationships between the top of the atmosphere long-wave outgoing radiation and BT (Wu & Yan, 2011; Yang & Slingo, 2001). We tested both approaches and found that they result in very similar estimates, especially over the convective regions, which are the focus of this study. Since empirical estimates are significantly cheaper to perform compared to running a radiative transfer model, we use the equation presented in Wu and Yan (2011) for the calculation of the BTs using WRF top of the atmosphere outgoing long-wave radiating. This approach is similar to what has been used in existing MCS studies (Feng, Leung, et al., 2021).

Since $Z$ is no prognostic model variable and therefore must be diagnosed from quantities simulated in the model microphysics scheme. In the P3 scheme $Z$ is directly calculated from integrating over the particle size distribution using the particle mass-size (m-D) relations predicted by the scheme (that evolve with time and for different particle sizes), with dielectric constants for published literature (Smith, 1984). In the Thompson and Morrison scheme, $Z$ is also calculated by integrating over the particle size distribution with the same dielectric constants, except they use fixed m-D relations for ice/snow/graupel and include a calculation of the bright band by assuming liquid coating of snow/graupel above freezing (Blahak, 2007). The bright band from partially melted ice is neglected in the version of P3 that we use, though there is a newer version that does this explicitly as it predicts the liquid fraction on mixed phase particles.

2.4. Model Evaluation

Evaluating the performance of a model for simulating single convective events is difficult since small shifts in the time and location of a simulated storm can result in large differences when using traditional model evaluation metrics such as root-mean-square errors. In weather forecasting, this is widely known as double penalty problem (Prein et al., 2013; Roberts & Lean, 2008). We introduce a method that addressed the double penalty problem by identifying the time, location, and rotation in the simulation that best aligns with the observed MCS overpass over the ARM sites. The model analysis is performed at the identified optimal, displaced location and rotation by using observations that are common to many regions where MCSs occur.

In Figure 2, we provide a schematic of our approach. It starts by selecting a scan area ($N$), which is based on the spatial reach of the SIPAM S-band radar in Manaus (∼180 km radius at 2 km height). A square box scan area with a side length of $N = 2.6°$ (∼290 km) was selected for computational efficiency. As previously described, $Z$ CAPPIs at 2, 4, and 6 km above ground level, are input, with the multiple heights included testing the sensitivity of the model evaluation to the height of the radar observation. Since BT and GridRAD observations are only available at full hours, we search for the time of maximum $Z$ in the scan area for MAO MCSs and round the time to the closest full hour.

Next, we define a search time window as the number of 10 min output intervals ($T$) that corresponds to the maximal allowed temporal displacement in the model. Here we use $T = ± 24$ output intervals (±4 hr) around the time of the observed MCS overpass. We decided to constrain $T$ to 4 hr since larger temporal displacements would likely result in MCSs that develop in significantly different environmental conditions than the observed MCSs. We used a similar rationale in defining a search area that is $2°$ larger than the scan area in each direction ($M = N + 2° + 2°$; Figure 2) and a rotational error $R$ that varies between ±30° in 5° intervals.

For each model output interval (t = 10-min), we derive the simulated $Z$ and BT within the search area, the search time window, and the rotational error window. For instance, we have 65 × 65 grid cells within the scan area and 165 × 165 grid cells within the search area, which results in 100 × 100 possibilities to shift the scan area within the search area. For each position we rotate the simulated filed between ±30°. For each output interval, we calculate two skill scores for all possible shifts of the scan area within the search area.

The first skill score is the spatial pattern correlation coefficient (CC; Wilks, 2011) which evaluates the model skill in capturing the spatial pattern of simulated $Z$ and BT without penalizing the model for systematic magnitude biases. The second skill score is the absolute cumulative distribution function difference (ACDFD; see Figure 2.
for an example). The ACDFD, in comparison, does not penalize the model for deficiencies in simulating spatial patterns, but solely focuses on the correct simulation of the \( Z \) and BT magnitudes. This evaluation results in a displacement matrix containing \( T \times (M - N) \times (M - N) \times R \) skill scores for \( Z \) CC, BT CC, \( Z \) ACDFD, and BT ACDFD. Using the above example, these two skill scores are calculated 6,240,000 times for each MCS case, assuming a search time window of ±4-hr with 10-min output intervals (48-time slices), 100 × 100 possibilities to shift the scan area within the search area, and 13 possible rotations. To speed up the calculation, we only calculate skill scores in 5-grid-cell intervals, which results in very similar outcomes but reduces the computational cost substantially (not show).

To find the temporal (\( \Delta t \)), spatial (\( \Delta x \)), and rotational (\( \Delta r \)) displacement that corresponds to the location of the optimal model performance, we combine these four skill scores by normalizing their distributions to a mean of zero and a standard deviation of one. Next, we multiply the normalized ACDFD matrices by minus one, which means that larger values are better (similarly to the CC statistics). The average of the four derived matrices is calculated, resulting in a displacement matrix in which larger values correspond to an improved agreement between the model and observations. Finally, we search for the maximum in the displacement matrix to derive \( \Delta t \), \( \Delta x \), and \( \Delta r \) and use the optimal location and time for model evaluation.

To better understand the impact of model physics on the MCS environments, we calculate CAPE, CIN, and vertical average hydrometeor mixing ratios in a ±15 grid cell square around the optimal location. CAPE and CIN are calculated with the python wrf.cape_2d function that finds the level of maximum equivalent potential temperature height in the lowest 3,000 m above ground. Next, a parcel with 500 m depth centered on this height is defined and used for the CAPE and CIN calculation.
3. Results

3.1. Idealized Tests

Before we apply the model evaluation method to our simulated MCSs, we test its performance on four idealized cases. These cases are similar to cases used in previous studies for testing model evaluation methods (e.g., see Figure 2 in Davis et al. [2009]) and exemplify how the derived skill scores are affected by specific biases in the simulation. To simplify the analysis, we remove the time dimension and only consider one variable (Z). A summary of these tests is provided in the list below.

- The first case represents a simulated MCS that is identical to the observed case, but eastward displaced by 4° (Figure 3a). The algorithm can detect the displacement, which is 350 km (4° longitude at 36.6° North; the latitude of the SGP site). As expected, the CC = 1, the ACDFD = 0 dBZ, and the rotation error is zero when accounting for the shift.

- The second case (Figure 3b) is identical to the first case, but the simulated Z values are 10 dBZ higher than the observed ones. The method can identify the displacement without any problems and returns a perfect correlation coefficient (CC = 1), zero rotational error, and an ACDFD of 6 dBZ. The reason why the latter is not 10 dBZ is because we are associating cloud-free areas with 0 dBZ, in both observation and simulations.

- The third test case (Figure 3c) features a displacement bias of 5° to the east and has a simulated MCSs width that is double the observed one. The algorithm detects an eastward displacement of 564.3 km (6.4°), a moderate CC of 0.56, a ACDFD error of 4 dBZ, and a rotational bias of 35° indicating that the MCS spatial patterns are erroneously simulated.

- The final test case (Figure 3d) features a simulated storm that is identical to the observed one but shifted by 7° to the east and rotated by 90°. The algorithm detects a eastward displacement of 587.9 km (∼7°), no bias in the ACDFD score, a rotational error of 90°, and a CC of 0.94. The reason for the non-perfect CC score are interpolation biases that occur during the rotation of the simulated field.

3.2. Evaluation of Simulated MCSs

After gaining confidence in our evaluation method based on idealized settings, we apply the evaluation algorithm to real-world MCS simulations over the SGP and MAO. Figure 4 shows a representative example of the algorithm's input and output for the 6 June 2014 case at 9:00 UTC at the SGP site. The observed MCS shows a large anvil cloud shield (Figure 4a) associated with a squall line (Figure 4g). The simulated cloud shield (Figure 4b) is smaller than the observed one although the model overestimates Z at 4 km altitude (Figure 4h). The location where the modeled MCS is most similar to the observed one is 180 km displacement toward the northeast (distance between red and blue dots in Figures 4a and 4b) and occurs 120 min later (Figure 4m). No rotational bias is detected. Comparisons between BTs in the observed (Figure 4c) and simulated (Figure 4d) scan areas show that simulated cloud tops are warmer (Figure 4f; ACDFD = 7.6 K), and the spatial pattern correlation is CC = 0.32 (Figure 4e). The spatial structures of Z in the scan areas are better simulated (Figures 4i and 4j) with a higher CC of 0.63 (Figure 4k); however, the simulated Z values are higher than observed (Figure 4l; ACDFD = 2.1 dBZ).

Figure 4m shows the spatial maximum values of the normalized ACDFD Z, ACDFD BT, Z CC, and BT CC scores for every 10-min model output within the search time window. Note that the normalized ACDFD values have been multiplied by minus one, which means that larger values are better for all skill scores. While BT CC is the closest to the observations at t = 0 min, the maxima in the other skill scores are delayed. Equally weighting the four skill scores results in a temporal delay of 120 min between the observed and the simulated MCSs. Figure 4n shows the displacement matrix at t = 120-min with the maximum value being highlighted as a blue dot and rotational errors being shown as gray triangles. The displacement matrix component from the four skill scores are shown in Figures 4o–4r, each showing a displacement of the simulated MCS toward the northeast.

3.3. Model Physics Sensitivities

Figure 5 shows observational and simulated results for the 17 November 2014, MCS event at the MAO site to exemplify the impacts of different microphysics and PBL schemes on the simulated cloud and Z fields. The simulated fields are shown at the time of the optimal comparison to the observations. This event features a line of clouds produced by a weakly forced line of convection (Figure 5a). Somewhat unexpectedly for a tropical
MCS event, most of the tested simulations can capture the basic characteristics of this case. Clear outliers are the simulations that use the MYNN2.5 PBL scheme since they are the only simulations that are not able to capture the north-east to south-west-oriented line of cumulus clouds that are visible in the satellite observations. Instead, simulations using the MYNN2.5 scheme produces wide-spread, disorganized clouds that seem to be strongly influenced by the Amazon River (especially visible in the simulations using the Thompson and Morrison micro-physics schemes), which is not evident in the observations.

Figure 6 shows a “heat map” that provides an overview of the four skill scores including the spatial, temporal, and rotational displacements for all tested physics combinations (MCS cases for ZCAPPIs at 2 km above the ground). The large case-to-case variability at both locations is prominent, which appears larger than the sensitivity to the

Figure 3. Ideal test case experiment showing the impact of (a) displacements, (b) intensity, (c) shape, and (d) rotational biases on the spatial displacement (Δx), correlation coefficient (CC), and absolute cumulative distribution differences (CDF) skill score between an observed (blue) and simulated (red) storm system. The red rectangle shows the algorithm’s search area (M = 14°), the hypothetical location of a radar site (here the ARM SGP site), the approximate reach of an S-Band radar (black) rectangle (4.4°), displacement matrix (gray contour lines), and the optimal displacement location (blue circle and black dashed rectangle).
selected physics (a more detailed analysis on this topic is presented below). Additionally, there is little correlation between skill scores. This implies that well simulated BT patterns (e.g., the SGP case on 18 June 2016) do not infer well simulated $Z$ patterns or ACDFD values. This lack of consistency is surprising and will later be analyzed in more detail. Another surprising result is that skill scores are similar for the MAO and SGP MCSs despite their
Figure 5. (a) Example of the observed and (b–j) simulated BT (gray contorts) and radar reflectively (inlet in lower left) at 2 km for the 17 November 2014, MCS case in the Amazon. Results using the Thompson, Morrison, and P3 microphysics scheme are shown top down and YSU, MYJ, and MYNN2.5 planetary boundary layer scheme from left to right.
different environmental conditions. Our initial expectation was that the SGP MCSs might be better simulated due to the involvement of mid-latitude disturbances that help to initiate and organize the systems. However, such a difference is not obvious besides there being a slightly smaller spatiotemporal displacements for SGP MCSs. The largest difference between MCSs at these two locations are the lower (better) ACDFD $Z$ scores over MAO, which is in part due to the stronger and larger frontally driven MCSs in the SGP (i.e., relative $Z$ differences might be more similar). Another noteworthy difference between the two regions is that SGP MCSs tend to feature a slight counterclockwise rotational bias (on average), while MAO MCSs have a slight clockwise bias. There is little systematic effect from evaluating $Z$ at different altitudes (not shown). We note that slightly higher average CC values are observed for the SGP events that used the 4 km CAPPIs. However, ACDFD scores for $Z$ are the worst (highest) at this altitude. We attribute these discrepancies to radar “bright band” signatures (observed $Z$ enhancement owing to aggregation and melting) in observed $Z$ factor expected near the melting layer, something that model microphysics are struggling to simulate and standard forward model-radar operators do not capture.

Averaging the skill scores over all cases (bottom row in each section) show that differences between physics schemes are typically not significant due to the large case-to-case variability.

This large case-to-case variability motivates us to average the skill scores of all cases and the 2, 4, and 6 km $Z$ heights to better isolate the impact of the model physics on the simulation performance (Figure 7). Limited

Figure 6. (a) Heatmap of the reflectivity (Z) correlation coefficient, (b) BT correlation coefficient, (c) Z ACDFD, (d) BT ACDFD, (e) temporal displacement, (f) spatial displacement, and (g) rotational error. Each panel shows results from the 9 different physics perturbations (columns) and case experiments (rows) in the Amazon (mao; top block) and US central Great Plains (sgp; bottom block). The average performance (mean over all cases) is highlighted in a black rectangle. Hatching in the average scores indicates significantly different results compared to the Thompson-YSU simulations using the Mann-Whitney U-test ($p = 0.1$). Shown are results for a 2 km radar scanning heights.

Figure 7. Average skill scores over all cases and reflectivity heights in Figure 6. Note that the color bar ranges are narrower compared to Figure 6. Hatching shows significant differences compared to the Thompson-YSU simulations using the Mann-Whitney U-test ($p = 0.1$).
consistency exists between schemes that perform well for the SGP and MAO events. For instance, simulations using the MYNN2.5 scheme have the highest BT CCs for SGP cases, yet feature the lowest CCs for MAO MCSs (Figure 7b). This result is in agreement with previous examples as in Figure 5. Interestingly, a lower skill score in simulating cloud top structures (Figure 7b) does not directly translate to a lower skill score in simulating \( Z \) patterns (Figure 7a). Figure 8 shows that this is particularly the case for CC that are frequently negative (e.g., in the case of P3-MYNN2.5 simulations at MAO). Correlation coefficients are higher when comparing \( Z \) and BT ACDFD scores particularly for SGP MCSs and for MAO cases that use the Thompson microphysics in combination with YSU or MYJ. While it is hard to tell how strong \( Z \) and BT fields are related in reality, it is concerning that there is so little or no consistency between them in the model.

A more rigorous assessment of the sources of variability in the SGP and MAO cases based on variance decomposition (Déqué et al., 2007; Prein et al., 2011) is shown in Figure 9. Case-to-case variability is the largest source of uncertainty contributing between one-third to two-thirds of the total variability (Figures 9a–9l). This large case-to-case variability is caused by limits in the predictability of single cases due to the chaotic nature of atmospheric convection and the impact of the large-scale environment on model performance. The variability stemming from the choice of microphysics or PBL scheme is comparatively small. The choice of the PBL scheme is most important (contributing ∼20% to the total variability) for the ACDFD score of BT in the Amazon (Figure 9g), while the microphysics scheme selection is most influential in simulating the \( Z \) ACDFD score in the SGP (Figure 9f). The sizes of the rings in Figure 9 indicate the total variability. Variability in the ACDFD scores of \( Z \) and the BT CC (Figures 9e and 9f) are smaller for cases in MAO, while the opposite is true for temporal and spatial displacements (Figures 9i–9l).

We repeated the variance decomposition by averaging over all cases within a region to better highlight the other sources of variability besides the case-to-case variability (Figures 9m–9x). This reveals major differences between modeling sensitivities of MAO and SGP MCS cases. For instance, the PBL scheme substantially impacts Amazon brightness temperature CC (30%; Figure 9o) and ACDFD scores (80%; Figure 9s). On the other hand, SGP \( Z \) ACDFD scores are very sensitive to the microphysics parameterization (60%; Figure 9r) and \( Z \) CCs change substantially with height at which \( Z \) is measured (60%; Figure 9n). Additionally, variability caused by the interactions between the PBL and microphysics schemes are much larger over the Amazon and contribute close to 50% regarding the BT CC, \( \Delta t \) and \( \Delta x \) scores. Concerning the total variability at the two locations, BT ACDFD score and the temporal and spatial displacement variabilities at the MAO site are substantially larger than those at the SGP site while the variability in \( Z \) ACDFD scores is much smaller for MAO MCSs.

To better understand the impact of the tested physics settings on the MCS simulations at the MAO and SGP sites, we calculate the evolution of CAPE, CIN, and mean cloud condensates at the corresponding optimal locations averaged over all cases (Figure 10). One of the most noticeable differences is that MAO pre-MCS environments
have lower CIN values when using the MYNN2.5 PBL scheme (Figure 10c). This indicates that using MYNN2.5 results in enhanced mixing at the top of the PBL and supports the development of wide-spread convection, such as seen in Figure 5. Similarly, CAPE values are typically smaller in pre-MCS environments when using MYNN2.5, although the differences are not as clear as for CIN (Figure 10a). The impact of model physics on CAPE and CIN is less systematic for SGP MCSs (Figures 10b and 10d) except for the post-MCS environmental CAPE, which is the lowest when using Thompson microphysics and the highest when using the Morrison scheme.

Using the MYNN2.5 scheme at the MAO site results in a local maximum of cloud condensates at ~3 km height in the pre-MCS environments and during the MCS overpass (Figures 10e and 10f), while using the YSU and MYJ schemes leads to a less pronounced peak that is at a lower altitude. Consistent with the above analysis, this indicates that the MYNN2.5 scheme produces a deeper and strongly mixed PBL. Such differences are not obvious for SGP MCS cases (Figures 10h–10j).

### 3.4. Best Performing Model Settings

In this section, we calculate the average skill scores ranks for each physics combination. This allows combining the six individual scores to a single average rank score that ranges from zero (best performing option for all tested physics) to one (worst-performing option), with 0.5 indicating the average performance. For more details, see Section 2.4.
Overall, the results plotted in Figure 11a shows that the Thompson microphysics in combination with MYNN2.5 and YSU as well as the Morrison-MYNN2.5 perform best for SGP MCSs on average. However, the differences in performance are not statistically significant except for the P3-MYNN2.5 scheme, which performs poorest overall. The good performance of simulations using the Thompson scheme agrees with (Feng et al., 2018), who showed that using the Thompson scheme outperforms simulated MCSs in the central US that use the Morrison scheme. The robust performance of the Thompson scheme might be related to it being more tuned to match observations compared to Morrison and P3. The Thompson scheme was frequently updated and re-tuned over time while the Morrison and P3 schemes did not change significantly except for bug-fixes. The average sensitivity to the PBL scheme is smaller compared to microphysics sensitivities, although individual scores such as the temporal displacement show a clear sensitivity to the PBL scheme. Our simulations indicate the Thompson scheme performs best in capturing the Z distribution (ACDFD score), while it performs below average concerning spatial displacements (except in combination with the MYNN2.5 scheme). It is important to notice that the average scores are typically not significantly different from each other mainly due to the large case-to-case variability.

As discussed before, the MAO MCS cases are more sensitive to the selection of the PBL scheme than SGP cases, although there is a clear effect of a microphysics-PBL scheme interaction as well (Figure 11b). As shown before,
using the MYNN2.5 scheme results in a sub-optimal performance independent of the microphysics parameterization. The main contributors to the poor performance are its deteriorated simulation of $Z_{ACDFD}$ scores and BT statistics. Overall, the best performance is achieved with the Thompson and MYJ scheme, followed by Morrison-YSU, and Thompson-YSU. Interestingly, a poor-performing physics combination in one region can perform much better in the other region and vice-versa. This can be seen for the spatial and temporal displacement scores when using Thompson-YSU. Similar to the SGP analysis, average differences are typically not significantly different.

From our input sensitivity testing, we find that three physics combinations perform above average independent of the $Z$ height that is used for the analysis. Those are all PBL combinations with the Thompson scheme and Morrison-YSU. If the performance of an individual score is more important than the average performance, other physics combinations might be more appropriate such as the P3-YSU combination, which results in the highest BT CCs.

In addition to the tests of CAPPI altitudes, we test the sensitivity of model performance to radar coverage/capture domain. These tests are only possible at the SGP site using NEXRAD GridRad product given the available radar networks (Figure 12). Overall, sensitivities are generally smaller than the physics sensitivity. This implies that the selection of well-performing physics options is not affected by this setting.

### 4. Summary, Discussion, and Conclusion

We present a new model evaluation method that allows us to differentiate spatiotemporal displacement biases from biases in the simulated structure and intensity of the MCSs. We evaluate the skill of kilometer-scale models in simulating MCSs using SGP and MAO radar and GOES satellite observations. We are particularly interested in the impact of the model microphysics and PBL scheme on the simulations of mid-latitude and tropical MCSs. The main results of this study are as follows.

- Kilometer-scale models equally well simulate continental tropical and mid-latitude MCSs in terms of the spatial structure and intensity of the convection and the cloud top field. However, spatial and temporal displacements tend to be smaller in mid-latitudes, likely due to the ability of the model to capture the large-scale forcing-driven convection.
• Model physics that work well in mid-latitudes do not necessarily work well in the tropics and vice versa. For instance, simulations using the MYNN2.5 PBL scheme best simulate the pattern correlations of cloud BT in SGP but perform worst in the Amazon basin. Although each tested model physics setup has strengths and weaknesses, we can identify model settings that perform above average in both environments, such as the Thompson and Morrison microphysics with the YSU PBL scheme or the Thompson scheme with the MYJ PBL scheme. Finding model setups that work well in different environments is critical for global kilometer-scale modeling efforts.

• SGP MCS simulations are most sensitive to changes in the microphysics, in agreement with previous studies (Stephan & Alexander, 2014). Amazon MCSs are more sensitive to the PBL scheme formulation. This can be understood by the much lower Baroclinic instability in the Amazon basin compared to the central US and the larger importance of PBL processes such as shallow convection and horizontal temperature and moisture gradients (e.g., land-river contrasts) whose simulation is directly affected by the PBL scheme.

• MCS case-to-case variability is the largest source of variability in our model performance evaluation. This large variability often overshadows sensitivities to model physics settings and complicates the evaluation of the effect of the latter on model performance. Future studies could use an ensemble-based approach (i.e., simulating multiple realizations of each event) or the piggybacking approach, which helps to separate physical impacts from natural variability (Grabowski, 2019) to reduce the impact of case-to-case variability on the results.

Figure 12. Similar to Figure 11 but showing the impact of increasing the scan area from 2.4° (left sub-columns) to 4° (right sub-columns) for US Great Plains cases. The radar scan height is 2 km.
• There is little correlation between the model’s performance in simulating Z and cloud BTs. BT characteristics are related to light ice and snow hydrometeors while Z characteristics are related to heavier precipitating hydrometeors such as rain and graupel/hail. This indicates model deficiencies in the creation and lifting of frozen hydrometeors and the relationship between in-cloud latent heating and anvil spread.

It is important to emphasize that the presented model performance is a combination of the predictability of the MCS events, the accuracy of the ERA5 forcing data, and errors in the WRF simulation. We did not investigate the impact of changing the initial and boundary forcing data on our results, but instead selected MCS cases that were well simulated (i.e., predictable) by downscaling ERA5 data. Also, we did not have the computational resources to run ensemble simulations for each case, which would have allowed us to investigate the role of predictability in our results. Instead, we averaged over an ensemble of deterministic runs, which in part accounts for the impact of predictability on the model results.

Future studies should focus on the simulation of the 3D structure of Z in kilometer-scale simulations to better understand potential biases in the vertical structure of MCSs due to deficiencies in simulating convective properties (e.g., up-and down-drafts as shown in Wang et al. [2020]) at 4 km grid spacing in combination with biases in the model physics. A high-priority research area is to better understand the ability of kilometer-scale models to simulate oceanic MCSs, particularly over the tropics due to their importance for the global water and energy cycle.

The results from this study will inform the model setup of additional MCS simulations in the US Southern Great Plains and the Amazon basin at different horizontal grid spacings. These simulations will assess the bulk and structural convergence of MCS characteristics in these two environments and will help to improve the representation of MCSs in weather and climate models.

Data Availability Statement

ERA-5 reanalysis data can be accessed from the Copernicus Climate Data Store (Copernicus, 2022). GOES13 and GOES15 satellite brightness temperature observations can be accessed from NOAA (NOAA, 2022). GRIDRAD data can be downloaded from Bowman and Homeyer (2017). Gridded GoAmazon2014/15 campaign SIPAM S-band radar data can be accessed from ARM (2022). The WRF model is open source and can be downloaded from WRF (2022). The model simulations can be accessed upon request from the corresponding author. The code that was used to analyze and visualize the presented data can be accessed at Prein (2022b). The WRF simulation output that is analyzed in this study can be accessed through Globus as described in Prein (2022a).

Acknowledgments

This research was supported by the U.S. Department of Energy Atmospheric System Research Program, an Office of Science Biological and Environmental Research Program (Grant No. DE-SC0020050). S. Giangrande and D. Wang of Brookhaven Science Associates, LLC. are supported under Contract DE-SC0012704 with the U.S. DOE. Data were obtained from the Atmospheric Radiation Measurement (ARM) user facility, a U.S. Department of Energy (DOE) Office of Science user facility managed by the Biological and Environmental Research Program. NCAR is sponsored by the National Science Foundation. We would like to acknowledge high-performance computing support from Cheyenne (https://doi.org/10.5065/D6RX99XZ) provided by NCAR’s Computational and Information Systems Laboratory, sponsored by the National Science Foundation under Cooperative Agreement No. 1852977. This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357.

References

Anselmo, E. M., Machado, L. A., Schumacher, C., & Kiladis, G. N. (2021). Amazonian mesoscale convective systems: Life cycle and propagation characteristics. International Journal of Climatology, 41(7), 3968–3981. https://doi.org/10.1002/joc.7053
Arm. (2022). GoAmazon2014/15 campaign SIPAM S-band radar [Dataset]. DOE Office of Science. Retrieved from https://www.arm.gov/data/berthou, S., Kendon, E. J., Chan, S. C., Ban, N., Leutwyler, D., Schär, C., & Fosser, G. (2020). Pan-European climate at convection-permitting scale: A model intercomparison study. Climate Dynamics, 55(1), 35–59. https://doi.org/10.1007/s00382-018-4114-6
Blahak, U. (2007). RADAR_MIE_LM and RADARN_AILIB—Calculation of radar reflectivity from model output (Technical Report). KIT Institute for Meteorology and Climate Research.
Bowman, K. P., & Homeyer, C. R. (2017). GridRad—Three-dimensional gridded NEXRAD WSR-88D radar data [Dataset]. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. Retrieved from https://rda.ucar.edu/datasets/ds841.0
Clark, P., Roberts, N., Lean, H., Ballard, S. P., & Charlton-Perez, C. (2016). Convection-permitting models: A step-change in rainfall forecasting. Meteorological Applications, 23(2), 165–181. https://doi.org/10.1002/met.1538
Coniglio, M. C., Correia, J., Jr., Marsh, P. T., & Kong, F. (2013). Verification of convection-allowing WRF model forecasts of the planetary boundary layer using sounding observations. Weather and Forecasting, 28(3), 842–862. https://doi.org/10.1175/waf-d-12-00103.1
Copernicus. (2022). ERA5 hourly data on single levels from 1979 to present [Dataset]. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). Retrieved from https://climate.copernicus.eu/climate-reanalysis
Davis, C. A., Brown, B. G., Bullock, R., & Halley-Gotway, J. (2009). The method for object-based diagnostic evaluation (MODE) applied to numerical forecasts from the 2005 NSSL/SPC Spring Program. Weather and Forecasting, 24(5), 1252–1267. https://doi.org/10.1175/2009waf222241.1
Déqué, M., Rowell, D., Lüthi, D., Giorgi, F., Christensen, J., Rockel, B., et al. (2007). An intercomparison of regional climate simulations for Europe: Assessing uncertainties in model projections. Climatic Change, 81(1), 53–70. https://doi.org/10.1007/s10584-006-9228-x
Demer, L. J., O’Brien, T. A., Rieger, D., Vogel, B., & Cooke, W. F. (2016). Are atmospheric updrafts a key to unlocking climate forcing and sensitivity? Atmospheric Chemistry and Physics, 16(20), 12983–12992. https://doi.org/10.5194/acp-16-12983-2016
Wang, D., Giangrande, S. E., Schiro, K. A., Jensen, M. P., & Houze, R. A., Jr. (2019). The characteristics of tropical and midlatitude meso-scale convective systems as revealed by radar wind profilers. *Journal of Geophysical Research: Atmospheres, 124*(8), 4601–4619. https://doi.org/10.1029/2018JD030087

Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences* (Vol. 100). Academic press.

WRF. (2022). Weather research and forecasting (WRF) [Software]. wrfmodel/WRF. Retrieved from https://github.com/wrf-model/WRF

Wu, X., & Yan, J. (2011). Estimating the outgoing longwave radiation from the FY-3B satellite visible infrared radiometer Channel 5 radiance observations. *Chinese Science Bulletin, 56*(32), 3480–3485. https://doi.org/10.1007/s11434-011-4686-6

Xue, L., Fan, J., Lebo, Z. J., Wu, W., Morrison, H., Grabowski, W. W., et al. (2017). Idealized simulations of a squall line from the MC3E field campaign applying three bin microphysics schemes: Dynamic and thermodynamic structure. *Monthly Weather Review, 145*(12), 4789–4812. https://doi.org/10.1175/mwr-d-16-0385.1

Yang, G.-Y., & Slingo, J. (2001). The diurnal cycle in the tropics. *Monthly Weather Review, 129*(4), 784–801. https://doi.org/10.1175/1520-0493(2001)129<0784:tdcitt>2.0.co;2