The June 2012 North American Derecho: A testbed for evaluating regional and global climate modeling systems at cloud-resolving scales

Weiran Liu[1], Paul Ullrich[1], Jianfeng Li[2], Colin M. Zarzycki[3], Peter Martin Caldwell[4], L. Ruby Leung[5], and Yun Qian[6]

[1]University of California Davis
[2]Pacific Northwest National Laboratory
[3]Pennsylvania State University
[4]Lawrence Livermore National Laboratory (DOE)
[5]PNNL
[6]Pacific Northwest National Laboratory (DOE)

November 24, 2022

Abstract

In this paper, we introduce a testbed for evaluating and comparing climate modeling systems at cloud resolving scales using hindcasts of the June 2012 North American derecho. The testbed is applied to two models: the regionally-refined Simple Cloud-Resolving E3SM Atmosphere Model (SCREAM) at horizontal resolutions ranging from 6.5 to 1.625 km and the Weather Research and Forecasting (WRF) model with 4 km grid spacing. We find the simulation results to be highly sensitive to the initial conditions, initialization time, and model configurations, with initial conditions from the Rapid Refresh (RAP) producing the best simulation. Significant improvement is identified in the SCREAM simulations as horizontal grid spacing is refined. While a propagation delay of approximately 2 hours is found in both models, SCREAM at 1.625 km simulates the observed bow echo structure of the derecho well and predicts strong surface gusts that exceed 30 m/s. In comparison, WRF hardly produces surface wind over 25 m/s, and the derecho wind gust in WRF is 42-46% lower than in SCREAM. Moreover, WRF has a lower bias in simulating cold clouds but overestimates the precipitation intensity. Both models well reproduce the observed outgoing longwave radiation spatial patterns (Pearson correlation > 0.88) while they simulate larger areas of composite radar reflectivity > 40 dBZ by up to 4 times and underestimate the precipitating area by ~ 70\% in WRF and 47\% in SCREAM compared to observations.

The June 2012 North American Derecho: A testbed for evaluating regional and global climate modeling systems at cloud-resolving scales

W. Liu[1], P.A. Ullrich[1], J. Li [2], C. Zarzycki [3], P. M. Caldwell[4], L. R. Leung [2], Y. Qian [2]

[1]Department of Land, Air, and Water Resources, University of California-Davis, Davis, CA, USA
[2]Pacific Northwest National Laboratory, Richland, WA, USA
[3]Department of Meteorology and Atmospheric Science, Pennsylvania State University, University Park, PA, USA
[4]Lawrence Livermore National Lab, Livermore, CA, USA

Key Points:

- A testbed of observational products, diagnostics, and metrics is constructed to evaluate hindcasts of the June 2012 North American derecho.
- Hindcast results are sensitive to initial conditions, initialization time, horizontal resolutions, and convective and microphysics schemes.
- The Simple Cloud-Resolving E3SM Atmosphere Model at 1.625km successfully reproduces severe surface gusts with wind speeds above 30 m/s.

Corresponding author: Weiran Liu, wraliu@ucdavis.edu

Abstract
In this paper, we introduce a testbed for evaluating and comparing climate modeling systems at cloud resolving scales using hindcasts of the June 2012 North American derecho. The testbed is applied to two models: the regionally-refined Simple Cloud-Resolving E3SM Atmosphere Model (SCREAM) at horizontal resolutions ranging from 6.5 to 1.625 km and the Weather Research and Forecasting (WRF) model with 4 km grid spacing. We find the simulation results to be highly sensitive to the initial conditions, initialization time, and model configurations, with initial conditions from the Rapid Refresh (RAP) producing the best simulation. Significant improvement is identified in the SCREAM simulations as horizontal grid spacing is refined. While a propagation delay of approximately 2 hours is found in both models, SCREAM at 1.625 km simulates the observed bow echo structure of the derecho well and predicts strong surface gusts that exceed 30 m/s. In comparison, WRF hardly produces surface wind over 25 m/s, and the derecho wind gust in WRF is 42-46% lower than in SCREAM. Moreover, WRF has a lower bias in simulating cold clouds but overestimates the precipitation intensity. Both models well reproduce the observed outgoing longwave radiation spatial patterns (Pearson correlation > 0.88) while they simulate larger areas of composite radar reflectivity > 40 dBZ by up to 4 times and underestimate the precipitating area by $\sim 70\%$ in WRF and 47% in SCREAM compared to observations.

Plain Language Summary

This paper describes a testbed for evaluating model performance on a particular high-impact weather event – the June 2012 North American derecho, a storm event associated with extreme winds and precipitation. The testbed is applied to evaluate the Simple Cloud-Resolving E3SM Atmosphere Model (SCREAM) and the Weather Research and Forecasting (WRF) model at resolutions that resolve cloud systems. The performance of both models is shown to be sensitive to the dataset used for model initialization. Finer grid resolution generally leads to better model performance. All simulations show a 2-hour delay in predicting the evolution of the derecho and produce more intense rainfall. SCREAM generates a more realistic convective front than WRF and produces stronger surface winds. The evaluation protocol can be used to better understand the credibility of model simulations of extreme events and guide model development.

1 Introduction

Climate modeling systems are among our best tools for understanding the climate system and future impacts of climate change (Kharin et al., 2007). In the pursuit of models of the highest quality, modeling groups cycle between developing new functionality, testing that functionality in isolation, integrating it into comprehensive modeling systems, and evaluating those combined systems. In the case of climate models, evaluation has generally focused on average behavior over large regions or long time periods (Gleckler et al., 2008; Eyring et al., 2019). However, this form of generalized analysis does not address whether climate models are able to simulate the most extreme and high-impact weather phenomena, such as extreme mesoscale convective systems (MCSs), with high fidelity.To this end, in this paper we propose one such extreme event testbed for evaluating climate modeling systems that operate at cloud resolving scales. The testbed focuses on historical simulation of a single event, in this case the June 2012 North American derecho and accompanying MCS, with the intention of providing a standard and comprehensive suite of metrics for model assessment and intercomparison.

MCSs are responsible for a variety of severe atmospheric hazards, such as flood-producing heavy rainfall events (Schumacher & Johnson, 2006; Stevenson & Schumacher, 2014; Hu et al., 2020a), lightning (Carey et al., 2005), and damaging winds (Bernardet & Cotton, 1998; Schoen & Ashley, 2011). Due to their important role in the hydrological cycle (Hu et al., 2020b) and land-atmosphere interactions (Hu et al., 2021), there is considerable demand for evaluating their representation in one model or as part of an intercomparison across different models (Van Weverberg et al., 2013; Schumacher & Clark, 2014; A. F. Prein et al., 2020; Feng, Song, et al., 2021; Na et al., 2022). However, most previous studies focus on the long-term climatological metrics averaged over multiple years, which does not take specific events into consideration (Demaria et al., 2011; Pinto et al., 2015). A few studies evaluated the performance of climate models in the hindcast of individual extreme MCS events qualitatively but without a uniform set of metrics (Toll et al., 2015; Grunzke & Evans, 2017). Even in the short-term, for case studies of extreme MCSs that last for 1-2 days, evaluations are mainly conducted through qualitative analysis of spatial patterns (Toll et al., 2015; Grunzke & Evans, 2017) rather than quantifying the model performance using a uniform set of metrics. Previous studies (Davis et al., 2006; N. Roberts, 2008; N. M. Roberts & Lean, 2008; Mittermaier & Roberts, 2010) proposed and discussed fractions skill score (FSS) as a variation of the Brier skill score to assess a common dataset that consisted of WRF model precipitation forecasts in geometric cases. However, as indicated in Davis et al. (2006), this skill score only considered the precipitation and was highly dependent upon the threshold values and the domain sizes. While the FSS provided a measure of the spatial accuracy of precipitation forecasts, additional techniques are needed to determine behaviors of other features to gain a comprehensive understanding of the convective systems.

It is challenging to simulate individual convective storms accurately, due to the need to adequately resolve complex physical interactions between dynamical and microphysical processes over a wide range of scales (Stensrud et al., 2013; Houze Jr, 2004; Weisman & Rotunno, 2004; A. F. Prein et al., 2015; Feng et al., 2018). Previous studies have demonstrated that the performance of MCS simulations is greatly influenced by a number of factors, such as horizontal grid spacing (Tao & Chern, 2017; Squitieri & Gallus, 2020), initial conditions (ICs) (Vié et al., 2011; Brousseau et al., 2016; Weyn & Durran, 2017), model configuration (Schumacher & Clark, 2014), and choice of parameterizations (Elliott et al., 2016; Wheatley et al., 2014; Feng et al., 2018). As a result, sensitivity tests and simulation ensembles are often carried out in MCS studies to determine optimal model configurations. However, different MCS tracking algorithms and evaluation criteria are employed in these studies (Fiolleau & Roca, 2013; Haberlie & Ashley, 2019; Feng, Leung, et al., 2021), leading to possible inconsistencies in the reported results and a lack of clarity regarding the strengths and weaknesses of various models. Storm evaluation is also subject to uncertainties due to the observations or reanalyses selected as reference datasets and the selected thresholds (Kolios & Feidas, 2010; Huang et al., 2018). Therefore, a comprehensive and robust evaluation process and a uniform suite of metrics and diagnostics are much needed to streamline the process and provide greater comparability across climate modeling studies for understanding MCS features and impacts, particularly in the context of large ensembles.

The testbed proposed herein can be used to evaluate the representation of multiple storm characteristics in regional and global climate models at cloud system resolving scales. The proposed evaluation protocol is subsequently applied to compare and contrast regional and regionally-refined global climate models for a specific severe storm event, which we recommend as a standard for broader intercomparison. In this study, we limit our investigation to the Weather Research and Forecasting (WRF) model at 4km and the regionally refined model (RRM) approach using the Simple Cloud-Resolving E3SM Atmosphere Model (SCREAM) with different model configurations. Sensitivity tests address RRM grid spacing (6.5 - 1.625 km), differences between hydrostatic and nonhydrostatic dynamical cores, low-resolution and high-resolution model configurations, initialization time, and data source for the ICs.

This paper is organized as follows: section 2 presents the proposed testbed, including a brief introduction of the severe weather event we selected, observations employed, and metrics selected; section 3 describes the SCREAM-RRM and WRF models in detail; section 4 evaluates the simulation results of both models; finally, section 5 provides a summary of our findings and conclusions.

2 The June 2012 North American Derecho Testbed

In this section we describe the proposed testbed, based on a 24-hour hindcast of the June 2012 North American derecho, and designed for evaluation and intercomparison of climate modeling systems at cloud resolving scales. The testbed consists of a simulation protocol, a set of observational products, and a comprehensive set of diagnostics and statistical metrics that leverage those observations. Section 2.1 provides a meteorological overview of the derecho and previous relevant studies. Section 2.2 describes the selected observational datasets. Section 2.3 presents four essential storm features that are examined in this framework, and section 2.4 explains the calculations of the metrics.

2.1 Meteorology

Johns and Hirt (1987) categorized derechos as meteorological events with severe wind gusts and precipitation lasting for several hours, in conjunction with a linear MCS. An extensive study (Corfidi et al., 2016) more recently defined a derecho as an event with: 1) convectively induced wind damage and/or gusts of > 25.7 m/s over an area with a major axis of 400 km, 2) geographically-consistent reports, and 3) presence of 3 or more reports of gusts > 33.4 m/s within the affected area. Among all historical derechos in North America, the June 2012 North American derecho (or June 2012 Mid-Atlantic and Midwest derecho) is one of the most infamous – a progressive derecho event that became one of the most destructive and fastest-moving derechos in US history.

The June 2012 North American derecho was characterized by an intense bow-echo MCS causing widespread severe wind damage across the upper Midwest and the Ohio River valley, as well as the mid-Atlantic states, during the afternoon and evening of 29 June and early morning of 30 June in 2012 (Shourd, 2017; Shourd & Kaplan, 2021). This particular event was selected because of the significant socioeconomically-hazardous impact and the high forecast difficulty. At initiation, a relatively small cluster of storm cells began to form as embryonic convection in eastern Iowa around 14:00 UTC on 29 June. Around 16:00 UTC, the small storm cluster began rapidly forming a well-defined MCS before crossing through Chicago, Illinois. Afterward, the MCS expanded into an asymmetric bow echo over Indiana as it accelerated southeastward at about 25 m/s slightly to the north of the frontal boundary. The MCS intensified further as it crossed Indiana and Ohio, transforming into a derecho MCS. The MCS continued along its destructive path until reaching the Atlantic coast of Virginia and Maryland about 06:00 UTC on 30 June. As estimated by the Storm Prediction Center (SPC), a damaging wind swath of about 1000 km in length resulted from this event, with over 800 wind damage reports during the 10-hour lifetime. Severe wind gust reports ranging between 25–33 m/s were widespread with peak gusts in excess of 40 m/s reported over eastern Indiana and western Ohio.

As indicated in Johns and Hirt (1987), progressive derechos are frequently challenging for operational meteorologists to forecast due to their weakly forced nature. The June 2012 North American derecho was underforecasted days and hours ahead of time, as well as throughout much of the duration of the storm. Most numerical weather prediction models showed no indication that any convective cells would develop, illustrating the forecast difficulty (Halverson, 2014; Guastini & Bosart, 2016; Schumacher & Rasmussen, 2020).

This forecast difficulty serves as the motivation for the following studies. Fierro et al. (2014) evaluated the short-term forecast (≤ 6 h) of the derecho event from the regional WRF model at 3 km resolution to compare two distinct data assimilation techniques. Shourd and Kaplan (2021) simulated the derecho using the WRF model in a nested domain with the inner domain at 2 km resolution and reproduced the super derecho. However, no quantitative evaluation metrics were used in these two analyses, resulting in no clear conclusions drawn as to the quality of the reproduction, especially when compared with other studies. Shepherd et al. (2021) performed an 11-member ensemble of convection-permitting regional simulations using WRF and tested the sensitivities to model configuration including microphysics parameterizations, lateral boundary conditions, start dates, and use of nudging. All 11 members had difficulty capturing the realistic evolution of the derecho, exhibiting a time delay (ranging from 2 - 8 hours) in simulating the derecho intensification and passage.

Previous studies that focused on simulating the June 2012 North American derecho (Fierro et al., 2014; Shourd, 2017; Shourd & Kaplan, 2021; Schumacher & Rasmussen, 2020) have emphasized the analysis and evaluation of composite radar reflectivity. Nevertheless, wind gusts are an integral component of the definition of derecho. In order to provide a more thorough evaluation of the event, our study has an additional emphasis on evaluating the wind speed, along with precipitation and composite radar reflectivity.

2.2 Observations

It is well known that precipitation products diverge considerably across regions, even in the regional means at daily to seasonal timescales, and particularly across in-situ, reanalysis and satellite products (Miao et al., 2015; Beck et al., 2017, 2019; Sadeghi et al., 2021). Our testbed requires a detailed comparison of hourly precipitation pattern and magnitude at fine horizontal resolution, where the differences between these products are particularly large. Therefore, three high-resolution gauge-based precipitation datasets are used to evaluate the simulated precipitation:

1. The National Centers for Environmental Prediction (NCEP) 4km Gridded Stage IV Data (Lin & Mitchell, 2005; Du, 2011), which is a merged ground-based and radar-derived hourly rainfall accumulation dataset from 140 radars and ~ 5500 gauges over the continental United State (CONUS). The NCEP Stage IV dataset provides highly accurate precipitation estimates and is, therefore, widely used as a reference for the evaluation of precipitation (Hong et al., 2004; AghaKouchak et al., 2011, 2012; Nelson et al., 2016; X. Zhang et al., 2018).
2. NASA Integrated Multi-satellite Retrievals for Global Precipitation Measurement (IMERG) V06B final run (Huffman et al., 2015), which intercalibrates, merges, and interpolates all estimates of the Global Precipitation Measurement (GPM) constellation, infrared (IR) estimates, gauge observations, and other potential sensors' data with a $0.1^{\circ} \times 0.1^{\circ}$ spatial resolution and 30 minute temporal resolution.
3. NOAA Climate Prediction Center Morphing technique (CMORPH) bias-corrected V1.0 (Joyce et al., 2004; Xie et al., 2017, 2019) – this 8 km resolution dataset produces 30 minute estimates of rainfall derived from passive microwave observations and extrapolates them backwards and forwards in time via spatial propagation information obtained from geostationary IR satellite data.

Following previous efforts (Beck et al., 2019; Feng et al., 2018), we use the NCEP Stage IV dataset as our primary precipitation reference, but also provide supplementary results from IMERG and CMORPH. While the NCEP Stage IV precipitation dataset is of high quality, it is available only over the US. IMERG and CMORPH are included to generalize our framework for testbed cases worldwide, considering their broader coverage. An intercomparison of different precipitation datasets is out of the scope of this paper.

Outgoing longwave radiation (OLR) is evaluated using the brightness temperature (Tb) from the NCEP half-hourly 4 km IR V1 dataset (Janowiak et al., 2017), which contains globally-merged geostationary satellites with parallax correction and viewing angle correction. Tb is converted to OLR following the empirical formulation provided by Yang and Slingo (2001).

For observations of radar reflectivity, we use the hourly three-dimensional high-resolution Next-Generation Radar (NEXRAD) (Bowman & Homeyer, 2017), which covers most of the contiguous U.S merged from 125 National Weather Service WSR-88D weather radars. The raw spatial resolution of NEXRAD is 0.02°x 0.02°and a vertical resolution of 1 km. Composite reflectivity (cREF) is calculated as the maximum reflectivity for each column and time step in both NEXRAD and the simulations.

For wind speed evaluation, we use station records from the National Weather Service Automated Surface Observation System (ASOS) (Nadolski, 1998). There are 90 ASOS stations in the analysis domain (76°-88°W, 36.5°-42°N), shown as black circles in Figure 1c. The temporal frequency of the ASOS record is 5 minutes, although several records are missing. Two wind-related parameters from the ASOS are used:

1. **Wind speeds:** ASOS stations measure wind direction and speed once every second using meteorological equipment at a height of 10 meters. Five-second wind direction and wind speed averages are computed from the 1-second measurements. These 5-second averages are rounded to the nearest knot and retained for 2 minutes. The resolution of the wind speed is 1 knot and converted from knots to m/s in all analyses of this study.

2. **Gust wind speeds:** The gust wind speeds represent the maximum five-second wind speed measured in each five–minute period when gust criteria are met (Nadolski, 1998). Gusts are rounded up to the nearest whole knot and converted from knots to m/s. Gust wind speed is not a standard parameter and only reported when:
   (a) Gust wind speed is at least 3 knots (1.54 m/s) above the current running 2-minute mean wind speed.
   (b) Gust wind speed exceeds the minimum five-second average in the last 10 minutes by at least 10 knots (5.14 m/s).
   (c) The current 2-minute average wind speed is at least 3 knots (1.54 m/s).

2.3 Storm Characteristics

To provide a near comprehensive evaluation of the relevant meteorological characteristics of the derecho, the proposed testbed focuses on four essential parameters: precipitation, cREF, OLR, and wind speed. We define three features based on the commonly used thresholds in the previous MCS analyses to locate and track the derecho: the cold cloud shield, the precipitation feature, and the cREF feature. The cold cloud shield is defined as the contiguous area with Tb lower than 241 K (Maddox, 1980; Feng et al., 2018; Feng, Leung, et al., 2021). Following the empirical formulation provided by Yang and Slingo (2001), this Tb threshold is instead applied to the OLR (which is output directly from the models), using a threshold of 163.44 W/m^2. The precipitation feature is defined as the contiguous area with precipitation rate higher than 1 mm/hour (Peters et al., 2009; Yuan & Houze, 2010; Feng et al., 2018). The cREF feature is defined as a continuous area with composite radar reflectivity greater than 40 dBZ (Dye et al., 1989; Zipser & Lutz, 1994; Haberlie & Ashley, 2019).

The latitude and longitude of the midpoint of a certain feature is calculated as the mean of the maximum and minimum of the latitude and longitude of the object. While the centroid of the feature polygon could have been similarly employed (Pinto et al., 2015; Davis et al., 2006), we observed similar results to those obtained via the simple midpoint.

Therefore, this study only uses the midpoint instead of the centroid because of the simplicity of computation. The features are further isolated and tracked using TempestExtremes (Ullrich & Zarzycki, 2017; Ullrich et al., 2021), as shown in the appendix. The area of an isolated feature is calculated as the sum of areas of grid points that are detected in the TempestExtremes.

Note that the definitions of MCSs and thresholds are diverse in the past studies (Schumacher & Johnson, 2005; Yuan & Houze, 2010). While we choose the most widely used thresholds, the involvement of thresholds and tracking algorithms would still induce a certain degree of uncertainty, as mentioned in section 1. Therefore, we will use metrics without incorporating the storm detection and tracking if possible besides the features described above.

2.4 Evaluation Metrics

Quantitative evaluation of the SCREAM and WRF experiments is performed via a variety of statistical techniques over the analysis region shown in Figure 1c. To facilitate comparison, the simulated variables are interpolated onto the coarse observation grid (i.e., 0.05° for OLR and cREF and 0.1° for precipitation). Here we use precipitation as an example, but similar calculations are applied to other variables (OLR and cREF).

The model bias is measured by the mean error (ME),

$$ME = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i), \tag{1}$$

where N is the total number of verification grid point, and p and o are the simulated and observed values, respectively. Mean absolute error (MAE), is calculated as

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|p_i - o_i|. \tag{2}$$

The root-mean-square of the error (RMSE) (Anthes, 1983) is defined as

$$RMSE = \left[\frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2\right]^{1/2}. \tag{3}$$

The pattern correlation between simulations and observations are represented by the Pearson product-moment correlation coefficient, calculated as

$$r = \frac{\sum_{i=1}^{N}(p_i - \overline{p})(o_i - \overline{o})}{\sqrt{\sum_{i=1}^{N}(p_i - \overline{p})^2}\sqrt{\sum_{i=1}^{N}(o_i - \overline{o})^2}}. \tag{4}$$

We choose the centered form, which measures the similarity of the pattern after removing the regional mean (Santer et al., 1993), because it provides additional information independent of the mean bias. We also use the Spearman rank correlation coefficient as a robust and resistant alternatives to the Pearson product-moment correlation coefficient. The Spearman correlation is simply the Pearson correlation coefficient computed using the ranks of the data,

$$r_s = 1 - \frac{6\sum_{i=1}^{N}D_i^2}{N(N^2-1)}, \tag{5}$$

where D_i is the difference in ranks between the ith pair of values.

It is important to stress that a particular simulation can exhibit a bias close to zero, along with poor correlation (e.g., the regionally-averaged precipitation rate is similar to the reference dataset but the precipitation patterns are distorted), or a high correlation, but with a high bias (e.g., a consistent spatial distribution of precipitation but with intensified rainfall rate relative to that of the reference dataset). As such, the conclusions derived from single metrics could be misleading, suggesting a need to incorporate multiple measures in such an analysis.

Our evaluation metrics also include two scores normally used in the assessment of accuracy of weather prediction. The first is the bias score (BS), which indicates whether the model over or under predicts the fractional areal coverage of precipitation for a certain threshold. On the other hand, the threat score (TS), ranging from 0 (worst) to 1 (best), is used to measure the skill of predicting the area of precipitation exceeding a certain intensity threshold. The BS and TS are defined as

$$BS = \frac{P}{O}, \tag{6}$$

and

$$TS = \frac{H}{P + O - H}, \tag{7}$$

where P and O are the number of grid points with values higher/lower (i.e., higher for precipitation and cREF; lower for OLR) than the threshold in the simulation and reference dataset, respectively. H is the number of grid points higher/lower than the threshold in both the simulation and the observation.

3 Models

3.1 The SCREAM Regional Refined Model

A series of RRM simulations are conducted using SCREAM (Caldwell et al., 2021; Liu et al., 2022), configured with a high-resolution (HR) grid located in the northeastern US, a low-resolution (LR) grid covering the remaining globe, and a transition area between them (Figure 1b). Figure 1a shows the SCREAM RRM grid in the global orthographic projection. The grid is based on the unstructured cubed-sphere finite-element grid with 4 Gauss-Lobatto-Legendre (GLL) nodes per element's edge (`np4`). Our LR grid uses 128 × 128 spectral elements on each face, denoted ne128, corresponding to a horizontal grid spacing of 0.25°(∼ 28 km). Using the offline software tool SQuadGen (Ullrich, 2014), three RRM grids were constructed using the same low base resolution (ne128) and various high resolutions: ne512 (6.5 km), ne1024 (3.25 km), and ne2048 (1.625 km). The HR portion of the grid is large enough to comprise the region where the derecho initiated, as well as its propagation path. While the derecho eventually migrated to the Atlantic in its decay phase, our analysis only focuses on processes on land where the damage occurred and, therefore, does not cover broad oceanic area in the HR portion.

The RRM approach has been validated in other models over many regions of interest (Zarzycki & Jablonowski, 2014; Sakaguchi et al., 2015, 2016; Rhoades et al., 2018; Wu et al., 2017; Xu et al., 2018) and demonstrated to be effective for regional climate studies at a reduced computational cost compared to uniform GCMs. For example, Zarzycki and Jablonowski (2014, 2015) demonstrated improved skill in simulating tropical cyclones in the Community Atmosphere Model with a refined mesh (0.25°) over the North Atlantic at multidecadal timescale. Huang and Ullrich (2017) reproduced the geographic patterns of 26-year historical precipitation climatology over the western US with the variable-resolution Community Earth System Model with a fine grid resolution of 0.25°. Two studies (Sakaguchi et al., 2015; Tang et al., 2019) demonstrated that RRM reproduced the seasonal precipitation of the high-resolution model over the CONUS.

However, previous RRM studies were performed with the highest horizontal resolution of around 0.25°and seasonal or longer timescales. Our study adopt the RRM approach in SCREAM with finer horizontal resolutions from 6.5 - 1.625 km and a timescale of 1-day. Since no optimal grid spacing has been identified for MCS simulations (Weisman

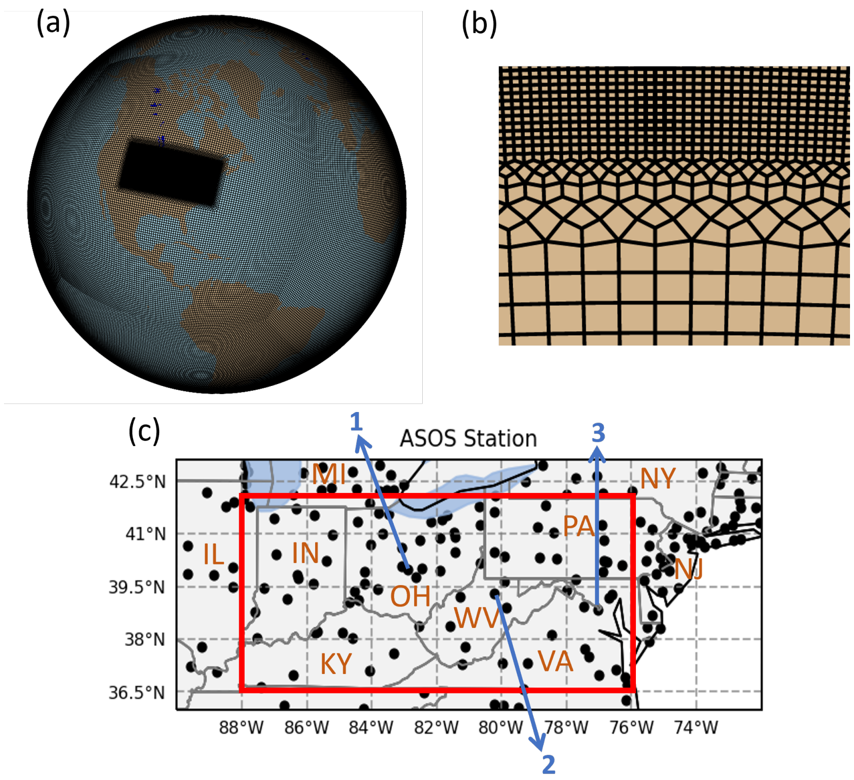

Figure 1. (a) The SCREAM RRM grid, shown using a global orthographic projection. (b) The transition region in the RRM grid from LR to HR resolution. (c) Locations of ASOS stations in black circles. The red box shows the analysis region (76°-88°W, 36.5°-42°N).

et al., 1997; Squitieri & Gallus, 2020), we investigate a variety of grid spacings between 6.5 km and 1.625 km to examine the impact of horizontal resolution on the derecho simulation in SCREAM.

Prescribed SST and sea ice extent are used for all SCREAM simulations. The land initial file is generated from a 12-month spinup land simulation prior to the initial date. Native output is saved every 15 minutes and later remapped using the TempestRemap software suite (Ullrich & Taylor, 2015; Ullrich et al., 2016) before calculating derived variables or performing analyses.

A summary of the SCREAM RRM simulations conducted in this study is provided in Table 1, including horizontal resolutions, ICs, initialization time, LR/HR configurations, and dynamical cores. Specifically, simulations SCREAM_6.5km, SCREAM_3.25km, and SCREAM_1.625km are designed to examine the sensitivity of model performance to grid spacing. Simulations SCREAM_ERA5 and SCREAM_ERAI serve as sensitivity tests of the model to the IC source. All simulations are initialized at 12:00 UTC 29 June 2012 and end at 12:00 UTC 30 June 2012, except for SCREAM_06Z initialized at 06:00 UTC 29 June 2012. The SCREAM HR configuration, where deep convection is turned off and includes a 128 layer vertical grid with a model top at 40 km (2.25 hPa), is employed in most simulations. SCREAM_LR uses the LR configuration; the model is run with 72 vertical levels with a top at 60 km, and the Zhang-McFarlane deep convection scheme (G. J. Zhang & McFarlane, 1995) is applied. These two configurations follow the vertical levels, model tops, and application of the deep convection scheme as described

Table 1. A summary of the SCREAM RRM simulations conducted and compared in this study.

Simulation Abbreviation	Fine Resolution	IC	Initialization Time (UTC)	LR/HR Configuration	Dynamical Core
SCREAM_6.5km	ne512 (6.5km)	ERA5+RAP	12:00 29 June 2012	HR	NH
SCREAM_3.25km	ne1024 (3.25km)	ERA5+RAP	12:00 29 June 2012	HR	NH
SCREAM_1.625km	ne2048 (1.625km)	ERA5+RAP	12:00 29 June 2012	HR	NH
SCREAM_ERA5	ne1024 (3.25km)	ERA5	12:00 29 June 2012	HR	NH
SCREAM_ERAI	ne1024 (3.25km)	ERAI	12:00 29 June 2012	HR	NH
SCREAM_06Z	ne1024 (3.25km)	ERA5+RAP	06:00 29 June 2012	HR	NH
SCREAM_LR	ne1024 (3.25km)	ERA5+RAP	12:00 29 June 2012	LR	NH
SCREAM_H	ne1024 (3.25km)	ERA5+RAP	12:00 29 June 2012	HR	H

Table 2. Timesteps of the SCREAM RRM simulations.

Simulation Name(s)	Fine Resolution	Dynamics Timestep (s)	Physics Timestep (s)
SCREAM_6.5km	ne512 (6.5km)	$16\frac{2}{3}$	300
SCREAM_3.25km, SCREAM_ERA5, SCREAM_ERAI, SCREAM_06Z SCREAM_LR, SCREAM_H	ne1024 (3.25km)	$8\frac{1}{3}$	100
SCREAM_1.625km	ne2048 (1.625km)	$4\frac{1}{6}$	50

in Caldwell et al. (2021) and Caldwell et al. (2019), respectively. The dynamical equations are solved using the High Order Method Modeling Environment (HOMME) (J. Dennis et al., 2005; J. M. Dennis et al., 2012; Evans et al., 2013). The simulations mostly use the HOMME nonhydrostatic (NH) spectral element dynamical core (Taylor et al., 2020; Bertagna et al., 2020; Liu et al., 2022) with one addition sensitivity test (SCREAM_H) utilizing the HOMME hydrostatic (H) dynamical core (Golaz et al., 2019; Caldwell et al., 2019; S. Zhang et al., 2020). Both the dynamics and physics timesteps are scaled across different RRM grids, controlled by the fine resolution, as shown in Table 2. Because of the horizontal resolution differences among three RRM grids, the topography is represented differently in these configurations.

The IC files are derived from three datasets: the Rapid Refresh (RAP) (Benjamin et al., 2016), the fifth generation of atmospheric reanalysis (ERA5) (Hersbach et al., 2018) and ERA-Interim (ERAI) (Dee et al., 2011), with details summarized in Table 3. A hindcast initialization suite (Betacast, Zarzycki and Jablonowski (2015)) is used to generate the IC files for the model from the above datasets. Since the RRM is a global model, ERA5 and ERAI data are directly mapped from the reanalysis grids to the model grid. The RAP analysis only covers North America, so for the simulations initialized using RAP, a two step approach is applied where a global 'base' IC is first generated using ERA5 and then the RAP analysis is used to overwrite the model state fields over the valid RAP region, displayed as ERA5 + RAP in Table 1. To eliminate noise associated with mapping the analyses across different grids, a hydrostatic correction is applied at each grid point to correct the hydrostatic surface pressure field between the analysis and model orographies, following the method described in Trenberth et al. (1993). Finally, all prognostic state variables in the vertical column are then reinterpolated based on the adjusted surface pressure since SCREAM uses a terrain-following coordinate.

3.2 WRF Model

WRF v4.3.3 (Skamarock et al., 2019) at 4 km is employed for intercomparison with the SCREAM RRM simulations. The WRF domain extends from 30.69°N to 48.04°N and from 102.78°W to 62.01°W. Four WRF simulations are run using different setups

Table 3. A summary of datasets used to generate IC files.

Dataset Name	Coverage	Temporal Resolution	Grid Spacing	Reference
RAP	North America	Hourly	13 km	Benjamin et al. (2016)
ERA5	Global	Hourly	0.25°	Hersbach et al. (2018)
ERAI	Global	6-hourly	0.75°	Dee et al. (2011)

Table 4. A summary of the WRF simulations.

Simulation Abbreviation	IC	Number of Vertical Levels	Microphysical Scheme
WRF_RAP	ERA5+RAP	45	Thompson
WRF_NARR	NARR	45	Thompson
WRF_HR	ERA5+RAP	72	Thompson
WRF_HR_P3	ERA5+RAP	72	P3

(Table 4) with the same simulation period, initialized on 12:00 UTC 29 June 2012, and output frequency as the SCREAM RRM simulations. The time step for integration is 10 seconds in the WRF simulations. The baseline simulation (WRF_RAP in Table 4), has 45 vertical layers with a thickness of $\sim$ 50 m for the lowest layer and a top at 100 hPa. Physics schemes used in WRF_RAP include the Thompson microphysics scheme (Thompson et al., 2008), the Rapid Radiative Transfer Model for General Circulation Models (RRTMG) shortwave and longwave radiation schemes (Iacono et al., 2008), the Mellor-Yamada-Janjic (MYJ) planetary boundary layer scheme (Janjić, 1994), the Eta similarity surface layer scheme, the Noah Land Surface Model (Chen & Dudhia, 2001), and the Building Energy Model coupled with the Building Environment Parameterization (BEP + BEM) for urban physics (Salamanca et al., 2010). Initial and lateral boundary conditions are from ERA5 and RAP, where ERA5 provides soil conditions while RAP provides atmospheric and land surface conditions.

Besides the baseline simulation, three sensitivity tests are performed (WRF_NARR, WRF_HR, and WRF_HR_P3; Table 4) to examine the impacts of different initial and boundary conditions, vertical resolutions, and microphysical schemes. The configuration of WRF_NARR is the same as WRF_RAP, except using initial and boundary conditions from the NCEP North American Regional Reanalysis (NARR) product (Mesinger et al., 2006). Compared to WRF_RAP, WRF_HR has 72 vertical layers with a vertical resolution of $\sim$20 m near the surface. The difference between WRF_HR and WRF_HR_P3 is that WRF_HR_P3 uses the Predicted Particle Property (P3) microphysics scheme with 3-moment ice (Morrison & Milbrandt, 2015).

4 Evaluation

Our discussion begins with a snapshot of the mature stage of the derecho at 00:00 UTC 30 June 2012 in section 4.1. The temporal evolution of the derecho is investigated in section 4.2 followed by section 4.3, which presents the metrics to quantify the fidelity of the models. We also display how to interpret the quantified metrics to understand the derecho characteristics in section 4.3. Section 4.4 provides additional discussion about the simulated 10-m wind speed.

4.1 00:00 UTC snapshot

Figures 2-3 show the instantaneous simulation outputs of OLR and cREF at 00:00 UTC 30 June 2012 in eight SCREAM RRM simulations (Table 1), two WRF simulations

(WRF_RAP and WRF_NARR, Table 4), and observations at 0.05° resolution. The observation panel is marked with red title in all figures. Unlike Figures 2-3, which show instantaneous outputs, Figure 4 shows the precipitation amount in the simulations and reference datasets accumulated from 00:00 to 01:00 UTC (since the NCEP Stage IV precipitation dataset is accumulated hourly). Some spatial displacement is clear between the cREF and precipitation due to the propagating nature of the derecho. We will discuss the two WRF sensitivity tests (WRF_HR and WRF_HR_P3; see Table 4) in section 4.4 and, therefore, not display their results in this section.

Figures 4a-c clearly show that the precipitation patterns in different products are divergent. While all three products include gauge corrections, IMERG shows significantly higher rainfall rates than NCEP Stage IV dataset and CMORPH, especially in the southern part of the derecho near (80°W, 39°N). However, the NCEP Stage IV rainfall is the most widely used reference dataset (Beck et al., 2019; Feng et al., 2018) and has the best agreement with the ASOS station records among all three products (not shown). Accordingly, NCEP Stage IV dataset is used as the primary precipitation reference dataset in the following analysis.

Based on the comparison of SCREAM RRM simulations at three different grid spacings (SCREAM_6.5km, SCREAM_3.25km, and SCREAM_1.625km) with observations, it is clear that simulations at higher horizontal resolutions appear to better represent the derecho. Specifically, the simulated derecho at 6.5 km resolution is underdeveloped, producing the smallest cold cloud shield (by $\sim 65\%$) and most compact cREF/precipitation feature. While the derechos at all three horizontal resolutions are all located upstream (northwest side) of the observed feature, the discrepancy between the simulation and the observation decreases as the resolution becomes finer. The bow-shape echo and the axis angle of the convective core are more qualitatively similar to observations in the 1.625 km simulation.

The simulation performance exhibits substantial sensitivity to the IC sources (SCREAM_3.25km, SCREAM_ERA5, and SCREAM_ERAI). This dependency has been pointed out in past research examining convection simulations, as summarized in section 1. Consistent with the results in Shepherd et al. (2021), despite the higher resolution and larger data assimilation volume of ERA5, the simulation initialized with ERA5 does not show significantly better performance than the one with ERAI. However, simulation initialization with RAP shows significantly improved performance compared to both ERA5 and ERAI. Notably, Figurski et al. (2017) also showed that simulations using ERA5 produce scattered reflectivity fields that are very different from those observed. WRF simulations (WRF_RAP and WRF_NARR) are also sensitive to the IC source, with better performance apparent in WRF_RAP than WRF_NARR. Interestingly, despite the good performance of SCREAM_3.25km, the simulation initialized 6 hours earlier at 06:00 UTC (SCREAM_06UTC) shows little precipitation and cREF, along with a weaker cold cloud shield, indicating the high sensitivity to the IC source even when applying the same dataset at different initialization times (Figurski et al., 2017).

The SCREAM simulation with LR model configuration (SCREAM_LR) is not able to reproduce the derecho successfully: namely, convective clouds do not form when the deep convective scheme is active. This is perhaps unsurprising, as previous studies have indicated better simulation of individual convective events in a convection-permitting model without a convective parameterization scheme than those in GCMs and RCMs (A. F. Prein et al., 2015; Fosser et al., 2015). This suggests a significant benefit comes from resolving convection explicitly, as the use of a convective parameterization scheme leads to common errors such as misrepresentation of the diurnal cycle of convective precipitation (Dai et al., 1999; Brockhaus et al., 2008) and the underestimation of hourly precipitation intensity (A. Prein et al., 2013; Fosser et al., 2015; Ban et al., 2014; Gao et al., 2017).

Although Liu et al. (2022) demonstrated the discrepancy between nonhydrostatic and hydrostatic simulations is significant over certain hotspots in the seasonal simulation ensembles, the simulation with hydrostatic dynamical core (SCREAM_H) is not significantly different from its nonhydrostatic counterpart in the snapshots of this short-term hindcast, producing remarkably similar result to SCREAM_3.25km. This suggests that even a hydrostatic dynamical core can simulate MCSs with comparable fidelity to a nonhydrostatic dynamical core, even far into the classical nonhydrostatic regime, potentially because the physics parameterizations dominate the model behaviors.

The detailed structure of the derecho in observations is particularly well simulated in the SCREAM_1.625km. Specifically, the bow-shape echo of the cREF core is tilted in a northeast-southwest direction, forming a classic bow echo described in Fujita (1978). The precipitation feature (Figure 4f) displays a similar tilting shape along with a larger precipitating area and higher rainfall intensity in the northeast tail. A secondary cluster is found in the southwest tail with relatively low rainfall rate in the center part of the derecho. In contrast, the shapes of precipitating and cREF features in the WRF_RAP simulation are aligned in a more east-west direction with the most intense rainfall showing in the northwest part of the derecho. The meridional spread in WRF_RAP ($\sim 1.5°$) is about half that in the observation and SCREAM_1.625km.

Even in the simulations with relatively better representation of the derecho (e.g., SCREAM_3.25km, SCREAM_1.625km, and WRF_RAP), the simulated precipitation rate is higher than the observed (i.e., NCEP Stage IV). These moist biases are consistent with past studies, such as a study of daily WRF hindcasts of monsoon convections in Moker Jr et al. (2018). While observational precipitation bias may be a factor here, there is no evidence to suggest this is the case.

In Figures 2-4, it is obvious that some simulations (i.e, SCREAM_ERA5, SCREAM_ERAI, SCREAM_06UTC, SCREAM_LR, and WRF_NARR) are not able to capture the derecho accurately and are simply not comparable to the other simulations. Therefore, in the following discussions, we will only show results in the better simulations (i.e, SCREAM_6.5km, SCREAM_3.25km, SCREAM_1.625km,and WRF_RAP). Given the clear similarity of SCREAM_3.25km and SCREAM_H, the result from SCREAM_H will also not be displayed, except when there is a noteworthy result.

As mentioned in sections 1 and 2.3, our study emphasizes the assessment of the simulated 10-m wind speeds because of its relevance to storm damage (Shourd, 2017; Shepherd et al., 2021). Figure 5 shows the simulated 10-m wind speed maximum (m/s; shaded) in the period of 00:00 - 01:00 UTC 30 June 2012, calculated as the maximum of 15-minute instantaneous wind speed. The region of high wind speed in Figure 5 is wider/larger than the instantaneous gust front because it includes the wind swaths over an hour, capturing the movement of the derecho. The dot markers indicate the gust wind maximum (m/s; left panels) and wind speed maximum (m/s; right panels) calculated from 5-minute ASOS stations' records. To simplify the figure, only ASOS stations with gust reports are shown in the left panels and right panels only show stations with wind speed maximum higher than 5 m/s. All simulation results are shown at native grid points without regridding to minimize interpolation error. The ASOS stations have the caveat that they possibly do not capture the highest wind speed due to their limited spatial and temporal coverage. Note that the ASOS gust wind speed is generally higher than regular wind speed by 2-10 m/s (see section 2.2 for details).

Compared with ASOS, SCREAM RRM performs well in simulating the observed 10-m wind speed. The bow-shaped convective feature that produces extended swaths of damaging surface winds is one of the most important feature of the derecho, which is clearly shown in SCREAM_1.625km as a curved wind front, related to either a very strong rear-inflow jet or a strong downdraft (Fujita, 1978). It is obvious that the area with high wind speed at 3.25 and 1.625 km resolutions is significantly larger than that at 6.5 km res-

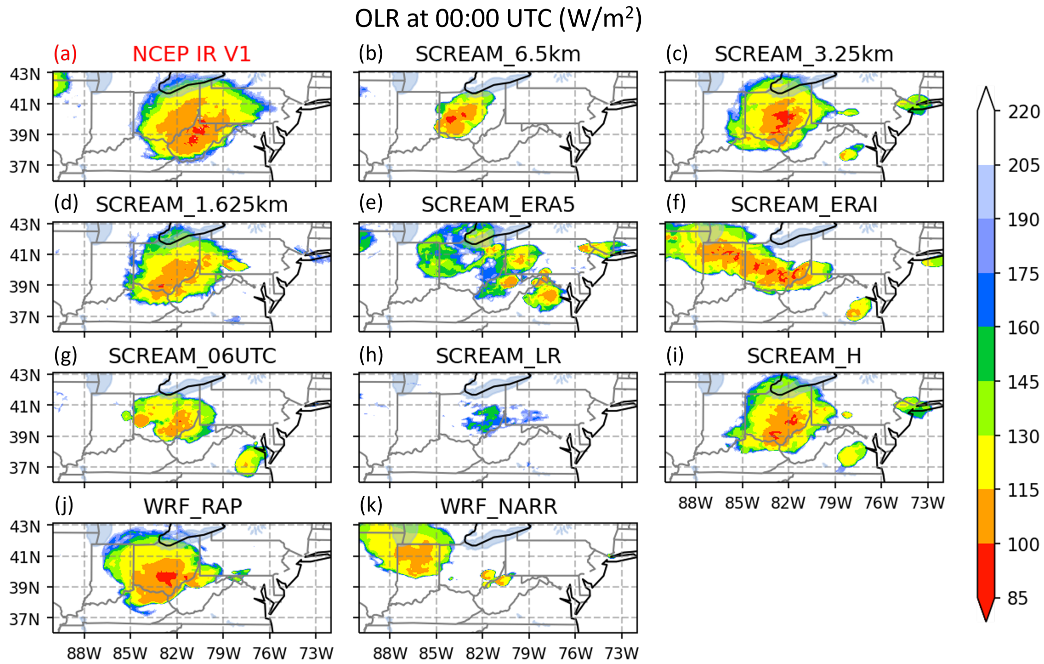

Figure 2. OLR (W/m^2) at 00:00 UTC 30 June 2012 in (a) NCEP IR V1, (b) SCREAM_6.5km, (c) SCREAM_3.25km, (d) SCREAM_1.625km, (e) SCREAM_ERA5, (f) SCREAM_ERAI, (g) SCREAM_06UTC, (h) SCREAM_LR, (i) SCREAM_H, (j) WRF_RAP, and (k) WRF_NARR. All datasets are remapped to 0.05° resolution. The panel with red title denotes the reference dataset.

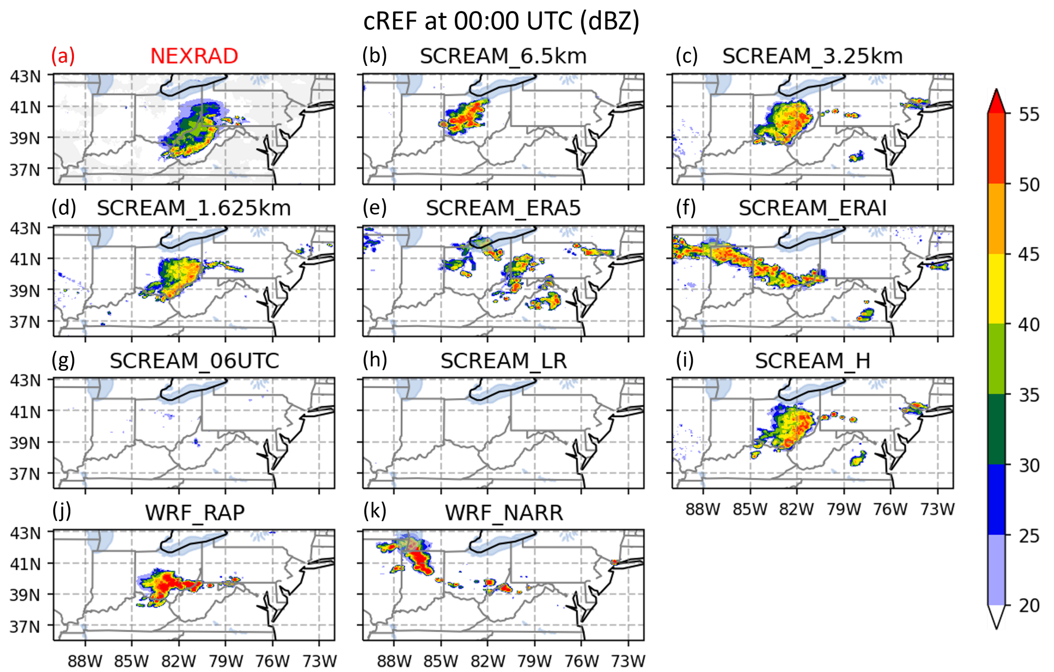

Figure 3. Same as Figure 2 but for cREF (dBZ). Panel (a) shows cREF in NEXRAD dataset.

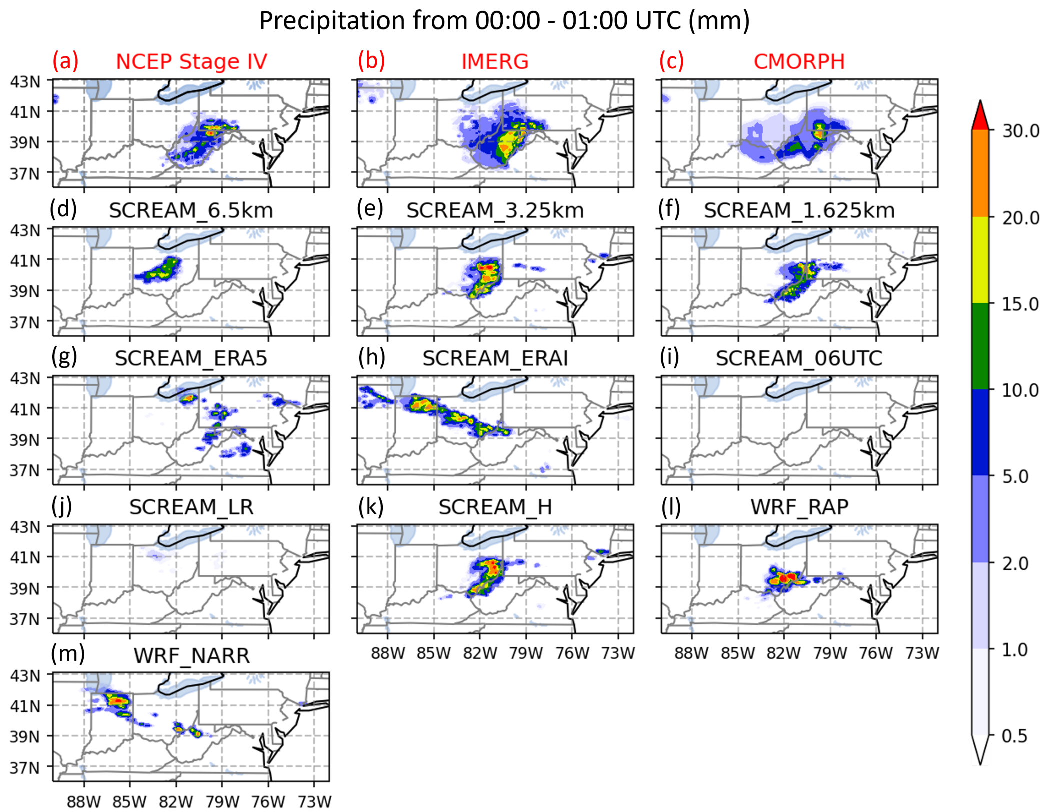

Figure 4. Same as Figure 2 but for accumulated precipitation (mm) from 00:00 - 01:00 UTC 30 June 2012. Panels (a-c) show NCEP Stage IV, IMERG, and CMORPH precipitation, respectively. All datasets are remapped to 0.1°.

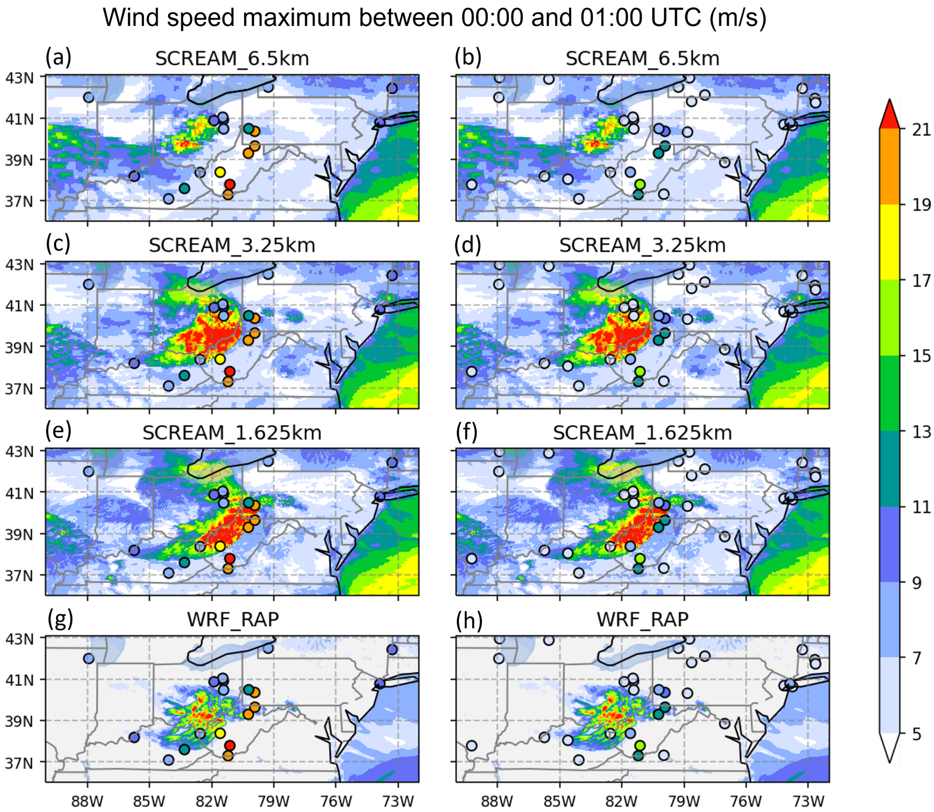

Figure 5. Wind speed maximum (m/s; shaded) between 00:00 and 01:00 UTC 30 June 2012 in (a-b) SCREAM_6.5km, (c-d) SCREAM_3.25km, (e-f) SCREAM_1.625km, and (g-h) WRF_RAP. The dot markers represent the ASOS gust wind speed maximum (m/s) in the left panels and wind speed maximum (m/s) in the right panels. The ASOS stations with wind speed maximum lower than 5 m/s are not shown in the right panels. All simulation results are displayed at raw grids.

olution, consistent with previous figures where enhanced fidelity is found at finer resolution. Additionally, the wind front forward of the derecho is closest to the ASOS sites with wind gust reports in the SCREAM_1.625km as the derecho location is simulated best at the 1.625 km resolution (Figure 4f). The WRF simulation shows lower wind speeds than SCREAM not only in the derecho-covered area but also over the entire analysis domain in general. More discussions about the simulated wind speeds and the different behaviors between WRF and SCREAM will be presented in section 4.4.

4.2 Time Evolution

Figure 6 shows 2-hourly evolution of cREF (dBZ) from 18:00 UTC 29 June to 06:00 UTC 30 June 2012 in the NEXRAD, SCREAM_6.5km, SCREAM_3.25km, SCREAM_1.625km, and WRF_RAP at 0.05° resolution. We only show cREF feature associated with the derecho identified using the TempestExtremes described in section 2.3 and remove other small clusters. Some time slots are not displayed (e.g., 18, 20, 22 UTC in SCREAM_6.5km) because the cREF feature does not qualified to be identified at that time (either too weak or too small using the defined thresholds).

The modelled track of the convective line broadly matches the observed one. The derecho-producing system proceeds southeastward from northern Indiana across central and southern Ohio with a strengthening convective core, reaching western West Virginia by 00:00 UTC. Over Ohio, the derecho system attains its greatest organization and strength. A rear-inflow notch at the back edge of system, which indicates an evaporatively cooled strong rear-inflow jet (Grim et al., 2009; Alliss & Hoffman, 2010), is evident before and during the leading line's transformation into a bow echo over Ohio. The mature bow echo contains two bookend vortices, generally marking a region of enhanced downdraft and an increased probability of stronger winds at the surface. The signature progressive bowing presentation is evident in the SCREAM simulation at 3.25 and 1.626 km resolutions. For a sufficiently persistent MCS, the Coriolis force eventually leads to a strengthening of the cyclonic (or poleward) bookend vortex and a weakening of the anticyclonic (or equatorward) vortex (Przybylinski, 1995; Schenkman & Xue, 2016). Accordingly, relatively fast eastward propagation is favored north of the front, with slower speed to its south in the observation and simulations. The storm system weakens as it moves into the southern New Jersey.

The development of the derecho from 18:00 to 22:00 UTC is significantly underestimated in the SCREAM_6.5km, which is corrected at finer resolutions. The observed weakening (displayed as a discontinuity in the track) around 04:00 UTC and the northward jump around 06:00 UTC near New Jersey are also well reproduced by the SCREAM_1.625km. The location of the derecho shows roughly 2-hour delay in SCREAM_3.25km and WRF_RAP, and larger ($\sim$ 3-hour) delay in SCREAM_6.5km. Comparing to SCREAM_3.25km, SCREAM at 1.625 km resolution reduces the delay by $\sim$ 0.5-1 hour. Longer postponements ranging from 3-8 hours were found in Shepherd et al. (2021), dependent on the model configurations.

All SCREAM and WRF simulations show larger cREF feature coverage than the NEXRAD. With that said, SCREAM_1.625km is the best ensemble simulating a narrow linear core with cREF $>$ 50 dBZ, most similar to NEXRAD, with extended spread in 40-50 dBZ. WRF shows significantly higher cREF than the SCREAM and NEXRAD by $\sim$ 10 dBZ, consistent with the overestimated rainfall intensity in Figure 4.

To compare the location of the simulated derecho to the observation accurately, Figure 7 shows the time series of longitude (left panels) and latitude (right panels) of the cold cloud shield (top), cREF feature (middle), and precipitation feature (bottom) from 18:00 UTC 29 June to 06:00 UTC 30 June 2012 in SCREAM_6.5km (dark blue), SCREAM_3.25km (yellow), SCREAM_1.625km (red), WRF_RAP (green), and observation (black). The solid line represents the center of the derecho at 15-minute frequency for all simulations and hourly frequency for the observation. Circle and triangle markers denote the 2-hourly maximum and minimum of the longitude/latitude, respectively.

All simulations show western and southern biases ranging from 0-3°, associated with the time delay of the migration. SCREAM_1.625km provides the best simulated position in the zonal direction among all simulations with the eastern progressive edge of the derecho following the observed one. WRF_RAP simulates the best location in the meridional direction while SCREAM_1.625km exhibits a more northern position by 0-0.5°.

Figure 8 shows the time series of the cold cloud shield, cREF feature, and precipitation feature areas. The dashed lines represent the raw simulation result frequency (15 minutes) while results averaged to the observation frequency (hourly for cREF and precipitation, and half hourly for OLR) are shown in the solid lines. The observed temporal evolution of cold cloud shield is reproduced best by WRF_RAP with the largest cold cloud shield present around 03:00 UTC. WRF_RAP shows a similar cold cloud size to SCREAM_3.25km before 00:00 UTC and grows up to twice the size of SCREAM after 00:00 UTC. SCREAM_1.625km captures the extending cold cloud shield before 23:00 UTC

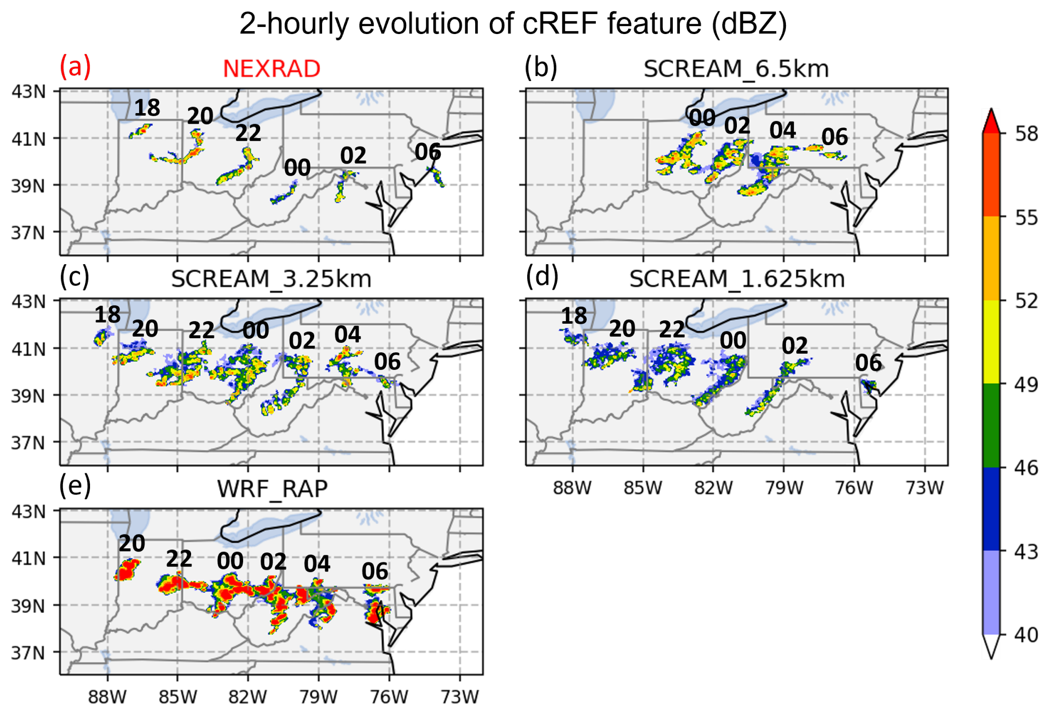

Figure 6. 2-hourly evolution of cREF feature (dBZ) from 18:00 UTC 29 June to 06:00 UTC 30 June 2012 in (a) NEXRAD, (b) SCREAM_6.5km, (c) SCREAM_3.25km, (d) SCREAM_1.625km, and (e) WRF_RAP at 0.05° resolution. The black bold numbers mark the hours in UTC.

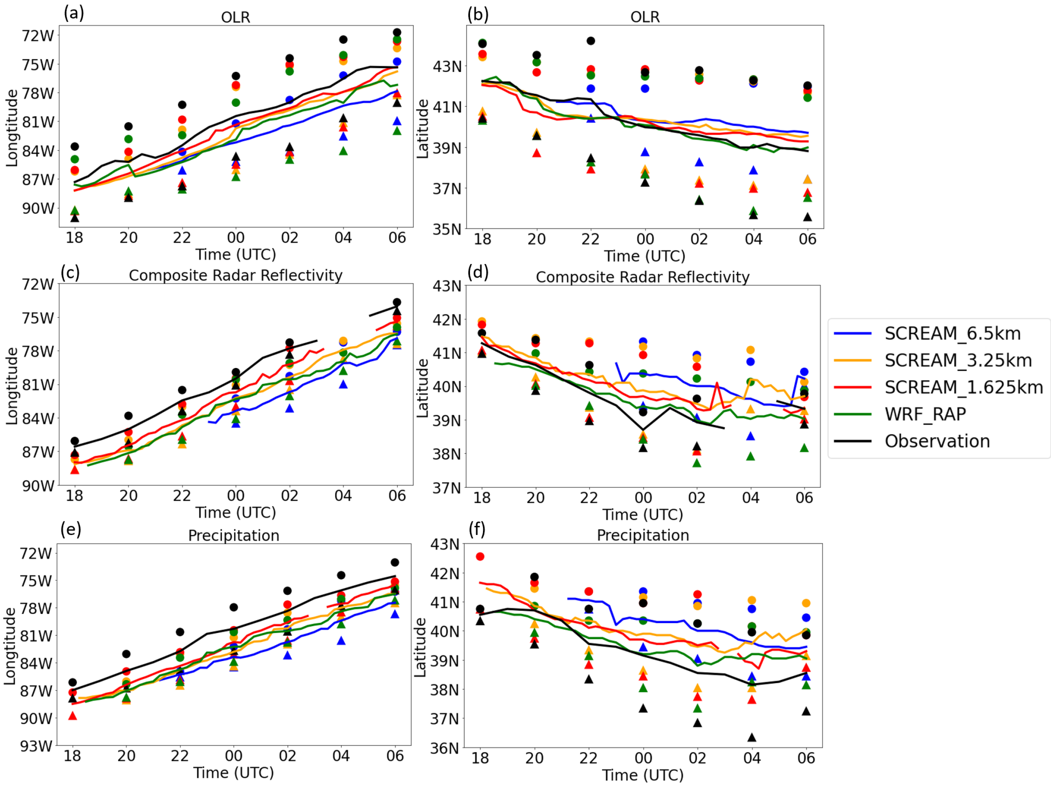

Figure 7. Time series of longitude (left) and latitude (right) of the cold cloud shield (top), cREF feature (middle), and precipitation features (bottom) from 18:00 UTC 29 June to 06:00 UTC 30 June 2012 in the SCREAM_6.5km (dark blue), SCREAM_3.25km (yellow), SCREAM_1.625km (red), WRF_RAP (green), and observation (black). The solid line represents the center of the derecho at 15-minute frequency for all simulations and hourly frequency for the observation. Circle and triangle markers denote the 2-hourly longitude/latitude maximum and minimum, respectively.

but shows a smaller cold cloud shield after 00:00 UTC by up to 50% than the observation.

All simulations overestimate the observed cREF feature size by up to four times while they underestimate the precipitation feature size by $\sim 50\%$. The precipitation feature is significantly larger than the cREF feature by up to 10 times in the observation (black lines in Figures 8b-c) indicating only approximately 10% of the precipitation feature is associated with high cREF (> 40 dBZ) whereas the rest of it has relatively low cREF. However, precipitation and cREF features show comparable sizes in WRF_RAP (greens lines in Figures 8b-c) suggesting almost the entire precipitation feature is associated with high cREF. In the SCREAM simulations, about 50% of the precipitation feature is associated with high cREF. The results are consistent with Figures 3 and 6 where the observation shows the most linear cREF area while WRF_RAP shows the widest coverage of the high cREF.

The observed cREF feature develops strongly between 18:00-20:00 UTC reaching its maximum coverage at 20:00 UTC, and persists until around 23:00 UTC. The precipitating area keeps expanding until 00:00 UTC when it starts to shrink while the cold cloud shield remains extending for three more hours. Similarly, in SCREAM simulations, the cREF and precipitation features show their coverage maxima 1-2 hours earlier than the cold cloud. Despite the propagation delay (Figure 6), the largest precipitating area of SCREAM occurs two hours earlier than observed, associated with the early decay of the cold cloud shield (Figure 8a). The peak time of precipitation, cREF, and cold cloud feature is almost simultaneous in the WRF simulation with a delayed cREF/precipitation feature area maximum by approximately 2.5-3 hours than the observation.

To evaluate the precipitation intensity, Figure 9a shows time series of regional-averaged precipitation rate in the analysis domain (76°-88°W, 36.5°-42°N), shown as the red box in Figure 1c, in the simulations and NCEP Stage IV precipitation dataset. Figure 9b is similar to Figure 9a but averaged only over precipitating grid points with rainfall rate > 1 mm/day. Additionally, averaged precipitation over the derecho identified by TempestExtremes is also examined (not shown) and implies similar results to Figure 9b. It is not shown considering the calculation processes in Figure 9b are much simpler and achievable for the broad research community without additional steps using TempestExtremes.

The observed regional-averaged precipitation peak time is captured by WRF accurately in both Figures 9a and b, but the averaged precipitation magnitude over the precipitating grid points is approximately twice as great as the observed (Figure 9b). WRF has wet bias over the precipitating grids but the precipitating area (Figure 8c) is reduced resulting in the similar magnitudes of precipitation peaks in the regional means (Figure 9a). While the maximum of averaged precipitation over the precipitating grid points is similar at three SCREAM resolutions, the time delay in the peak time is greatest in SCREAM_6.5km and declines at higher resolutions. SCREAM RRM simulations also show higher precipitation peaks than the observation by roughly 45%.

Figure 9c shows frequency distribution of the hourly precipitation rates at all grid points within the analysis domain (76°-88°W, 36.5°-42°N) during 18:00 UTC 29 June to 06:00 UTC 30 June 2012 in solid lines. The dashed lines are the same as the solid lines except for applying a 2-hour forward shift in the simulations, resulting in a period of 20:00 UTC 29 June to 08:00 UTC 30 June 2012; however, the conclusions are not sensitive to the 2-hour shift. WRF_RAP shows the strongest wet bias, strongly overestimating the observed precipitation rates higher than ~ 30 mm/day. However, WRF displays significantly lower frequency of precipitation rates below 30 mm/day and higher frequency for precipitation above 50 mm/day, consistent with a smaller coverage of relatively shallow precipitation and an overwhelming intense precipitation core in Figure 4. The SCREAM RRM simulations also produce an excess of extremely high rainfall rates (> 350 mm/day) but show lower frequency for the rainfall rates between 100 and 250 mm/day. SCREAM

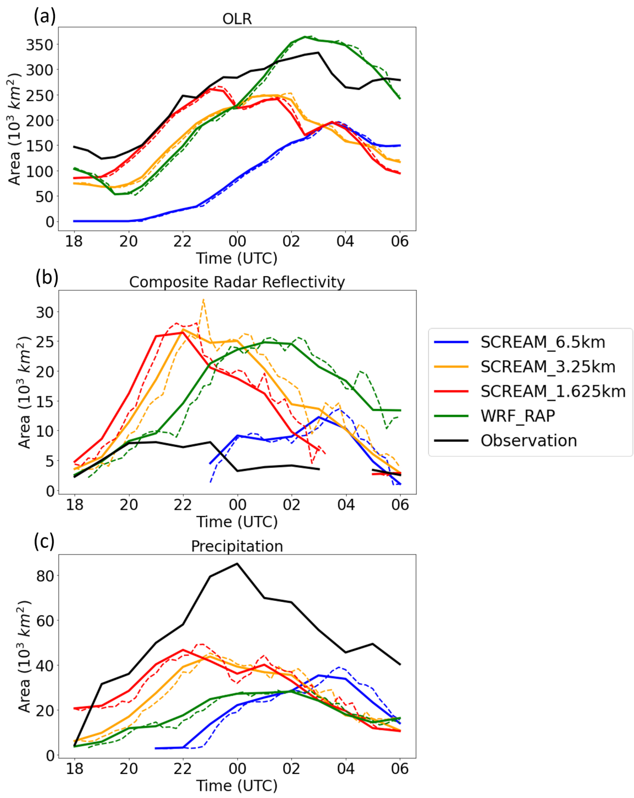

Figure 8. Time series of the area (10^3 km^2) of the identified features using (a) OLR, (b) cREF, and (c) precipitation from 18:00 UTC 29 June to 06:00 UTC 30 June 2012 in SCREAM_6.5km (dark blue), SCREAM_3.25km (yellow), SCREAM_1.625km (red), WRF_RAP (green), and observation (black). The dashed lines represent the results at 15-minute frequency for all simulations. The solid lines denote hourly frequency for cREF and precipitation and half hourly frequency for OLR.

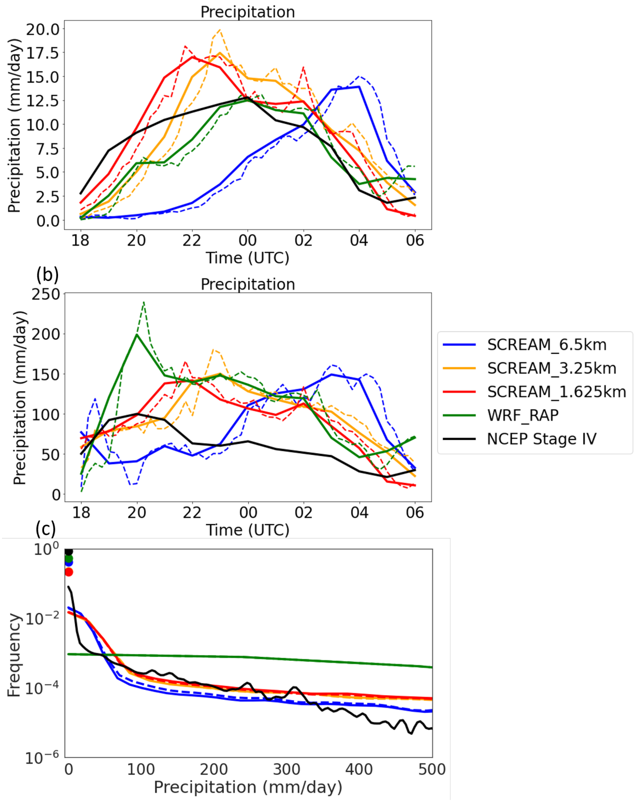

Figure 9. (a) Time series of regional-averaged precipitation rate (mm/day) in the analysis domain (76°-88°W, 36.5°-42°N), shown as the red box in the Figure 1c, from 18:00 UTC 29 June to 06:00 UTC 30 June 2012 in SCREAM_6.5km (dark blue), SCREAM_3.25km (yellow), SCREAM_1.625km (red), WRF_RAP (green), and NCEP Stage IV dataset (black). The dashed lines represent the simulation results at 15-minute frequency and the solid lines represent the hourly frequency. (b) is the same (a) but averaged only over precipitating grid points with rainfall rate > 1 mm/day. (c) Frequency distribution of hourly precipitation rates of all grid points within the analysis domain in the period of 18:00 UTC 29 June to 06:00 UTC 30 June 2012 in the solid lines. The dashed lines are the same as the solid lines but apply a 2-hour forward shift for the simulations resulting in a period as 20:00 UTC 29 June to 08:00 UTC 30 June 2012. The dots on the y axis denote the frequencies of zero precipitation.

Table 5. Locations of three ASOS stations.

Number in Figure 1c	Station ID	Station Full Name	State	Longitude	Latitude
1	KCMH	COLUMBUS PORT COLUMBUS INTL AP	OH	39.99139°N	82.88083°W
2	KCKB	CLARKSBURG BENEDUM AP	WV	39.29556°N	80.22889°W
3	KDCA	WASHINGTON REAGAN AP	VA	38.8483°N	77.0341°W

Table 6. Metrics derived from the OLR averaged from 18:00 UTC 29 June to 06:00 UTC 30 June 2012 in the analysis region (76°-88°W, 36.5°-42°N) in each simulation using NCEP IR V1 dataset as the reference. The calculations of the metrics are present in section 2.4. Scores in parentheses are calculated by applying a two-hour forward shift to the simulation results (i.e., the averaging period for the simulations changes to 20:00 UTC 29 June to 08:00 UTC 30 June 2012). The red numbers denote the best scores in each category. BS and TS are calculated using the threshold of 230 W/m^2.

Simulation Name	SCREAM_6.5km	SCREAM_3.25km	SCREAM_1.625km	SCREAM_H	WRF_RAP
RMSE	47.77 (39.35)	29.13 (26.43)	27.21 (25.29)	29.39 (27.69)	20.48 (17.65)
MAE	43.81 (34.61)	26.18 (22.61)	23.98 (21.46)	26.16 (23.89)	16.70 (14.15)
ME	43.81 (34.49)	22.31 (17.85)	18.86 (15.95)	21.66 (18.09)	8.18 (-0.11)
Pearson Correlation	0.89 (0.88)	0.89 (0.88)	0.88 (0.88)	0.87 (0.86)	0.88 (0.90)
Spearman Correlation	0.90 (0.86)	0.85 (0.80)	0.78 (0.78)	0.79 (0.74)	0.84 (0.82)
BS	0.20 (0.39)	0.58 (0.71)	0.73 (0.79)	0.61 (0.70)	0.83 (1.05)
TS	0.20 (0.39)	0.54 (0.64)	0.65 (0.71)	0.55 (0.61)	0.75 (0.87)

at 6.5 km exhibits a lower frequency in rainfall rates above 30 mm/day than simulations at 3.25 and 1.625 km, associated with the smaller precipitation feature (Figure 8c).

Figure 10 shows the time series of wind speeds at three ASOS stations, marked by the blue arrows in Figure 1c with details in Table 5. The three stations are selected to be airport stations in the derecho propagation path, spread over three states to capture various stages of the derecho life cycle, and using Figure 6 as a reference. In addition, the three stations are confirmed to not have missing wind speed records during the analysis period (18:00-06:00 UTC). The simulation results are shown at the closest single grid point to the specific ASOS station. While displaying time series at all ASOS stations is not feasible on a single plot, the three stations selected provide insights into the timing of the simulated gust fronts. Note that the ASOS station records have a high time frequency of 5 minutes and the simulation results are derived at individual grid points, causing the high-frequency fluctuations in the time series.

The delayed wind speed peaks representing the gust fronts are found in all simulations with reduced timing biases at finer resolution, consistent with the observed improvement in timing at 1.625 km resolution (Figure 6). The timing biases are approximately 1-1.5 hours at 1.625 km resolution, 2-3 hours at 3.25 km resolution, and 3-4 hours at 6.5 km resolution. The magnitudes of wind speed peaks in SCREAM_1.625km and SCREAM_3.25km are either comparable to or larger than ASOS winds (black lines) by 0-30%, and lower than ASOS gust speed (purple lines) by $\sim$ 30%. On the other hand, WRF_RAP shows the wind speed peak lower than ASOS wind speed by 27-70% and ASOS gust by 56-85%.

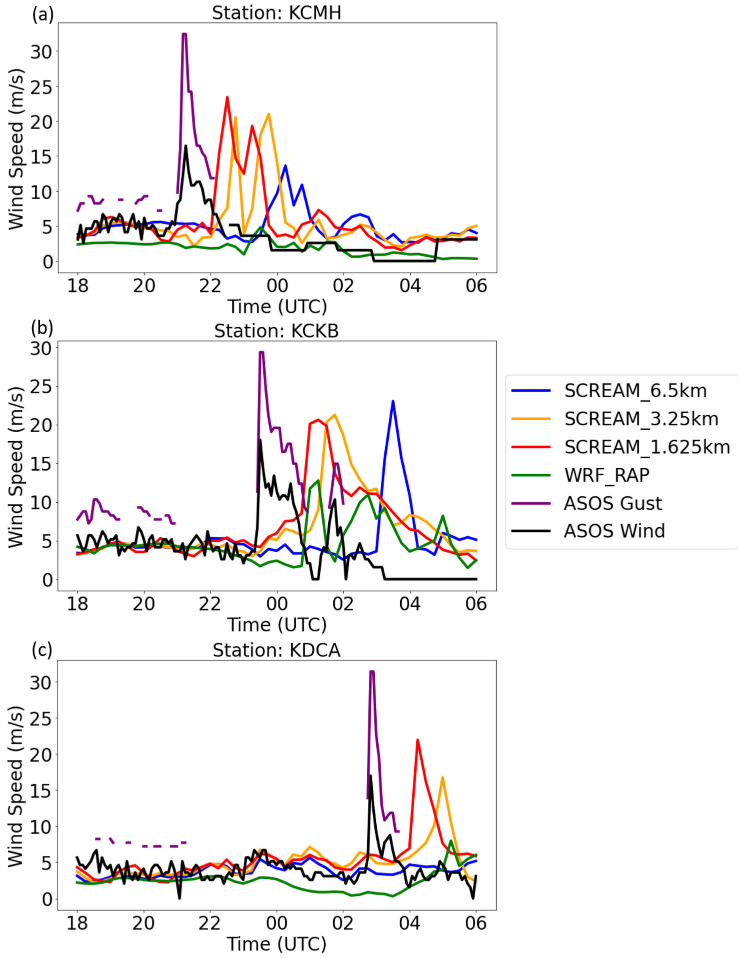

Figure 10. Time series of wind speeds (m/s) in the SCREAM_6.5km (dark blue), SCREAM_3.25km (yellow), SCREAM_1.625km (red), WRF_RAP (green), ASOS gust (purple), and ASOS wind (black) from 18:00 UTC 29 June to 06:00 UTC 30 June 2012 at ASOS station (a) KCKB, (b) KCMH, and (c) KDCA. The time intervals of ASOS records and simulation outputs are 5 minutes and 15 minutes, respectively. The simulation results are shown at the closest grid point to the specific ASOS station.

Table 7. As Table 6 but derived from the precipitation accumulated from 18:00 UTC 29 June to 06:00 UTC 30 June 2012 using NCEP Stage IV dataset as the reference. The last two columns show metrics calculated using the CMORPH and IMERG precipitation datasets comparing to the NCEP Stage IV dataset. BS and TS are calculated using the threshold of 15 mm.

Simulation/Observation Name	SCREAM_6.5km	SCREAM_3.25km	SCREAM_1.625km	SCREAM_H	WRF_RAP	CMORPH	IMERG
RMSE	6.99 (7.14)	8.50 (8.48)	7.73 (7.52)	8.17 (8.07)	9.26 (9.05)	5.78	8.65
MAE	4.32 (4.37)	4.66 (4.65)	4.49 (4.34)	4.81 (4.71)	4.87 (4.78)	3.88	5.47
ME	-1.73 (-1.46)	-0.70 (0.67)	1.03 (0.71)	1.16 (0.96)	-0.67 (-0.71)	2.89	4.56
Pearson Correlation	0.54 (0.53)	0.50 (0.51)	0.54 (0.55)	0.54 (0.54)	0.52 (0.53)	0.77	0.72
Spearman Correlation	0.70 (0.72)	0.73 (0.75)	0.71 (0.73)	0.71 (0.73)	0.73 (0.74)	0.87	0.86
BS	1.01 (1.07)	1.46 (1.48)	1.49 (1.41)	1.72 (1.67)	1.07 (1.08)	2.17	2.54
TS	0.28 (0.27)	0.23 (0.23)	0.23 (0.24)	0.24 (0.25)	0.25 (0.25)	0.30	0.28

4.3 Metrics

Table 6 displays metrics derived from the OLR in each simulation (see section 2.4 for the calculations of the metrics) with NCEP IR V1 dataset as the reference. The red number marks the best score in each category. The scores in parentheses are calculated by applying two-hour forward shift to the simulation results, providing better results (smaller biases) in all metrics except for two correlation scores.

WRF_RAP produces smaller biases in OLR than SCREAM when compared to observations. RMSE, MAE, and ME are lowest for WRF_RAP, particularly in the two-hour shifted ones, indicating WRF_RAP simulates the OLR field better than SCREAM. It is notable that SCREAM at finer resolutions shows better performance (in RMSE, MAE, and ME) than at coarser resolutions. The positive-biased OLR indicates a lower cloud top along with a smaller cold cloud shield (Figure 8a) in the simulations than the observed. The differences of Pearson correlations among all simulations are minor (< 0.03). Interestingly, despite previous analyses and other metrics showing better performance at finer resolution, Spearman correlation is highest in the coarsest simulation (SCREAM_6.5km), possibly caused by the underestimation of the cold cloud area at finer resolutions after 23:00 UTC (Figure 8a). BS showing values < 1 also indicates an underestimated cold cloud area in the simulations. WRF_RAP has the best representation of the cold cloud shield area, as indicated by the BS closest to 1, especially in the two-hour shifted BS.

The simulation using the H dynamical core (SCREAM_H) is also listed in Table 6. While it is not significantly different from the NH simulation in the snapshots of section 4.1 (Figures 2-5) and the time series in section 4.2 (not shown), we investigate whether the difference is more pronounced as the period is prolonged here. The H simulation shows slightly higher biases ($< 6\%$) in the two-hour shifted RMSE, MAE, and ME than SCREAM_3.25km, but the difference is much smaller than that among other simulation ensembles. As such, we attribute this difference to simulation variability rather than structural uncertainty.

Table 7 is the same as Table 6 but derived from precipitation accumulated from 18:00 UTC 29 June to 06:00 UTC 30 June 2012. Figure 11 shows the accumulated precipitation patterns along with two-hour shifted patterns shown in Figure S1. WRF_RAP shows the largest biases in RMSE and MAE, related to the overestimates of precipitation (Figures 4, 9, and 11). However, the ME in WRF_RAP is smallest in magnitude and becomes the best score among all simulations because the wet bias from enhanced precipitation intensity is offset by the reduced precipitating area (Figures 8c and 11e). The ME changes from negative to positive when SCREAM resolution becomes finer and the precipitating area of the derecho increases (Figure 8c). The ranges of RMSE and MAE in the simulations are comparable to those in CMORPH and IMERG, suggesting reasonable model performance in line with observational uncertainty. The Pearson and Spear-

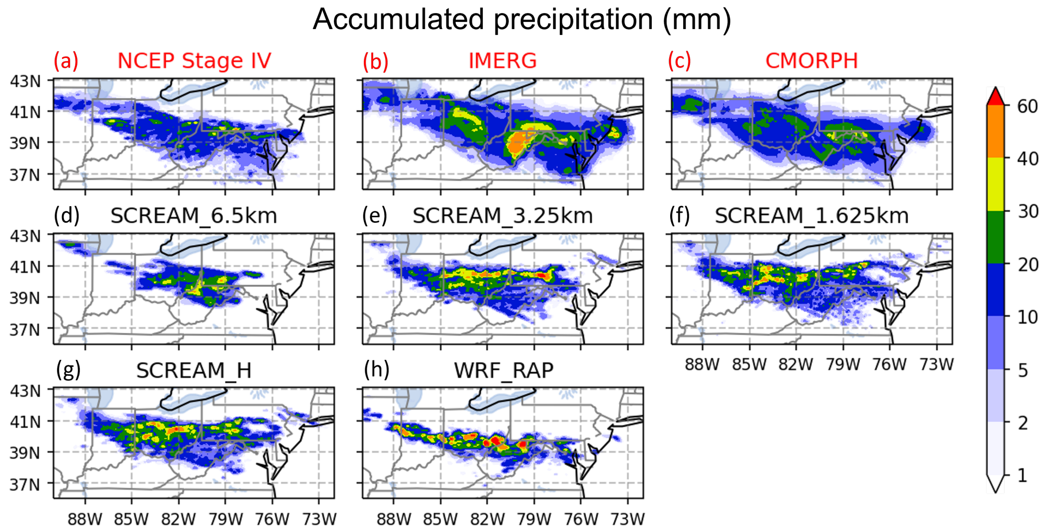

Figure 11. Accumulated precipitation (mm) from 18:00 UTC 29 June to 06:00 UTC 30 June 2012 in (a) NCEP Stage IV, (b) CMORPH, (c) IMERG, (d) SCREAM_6.5km, (e) SCREAM_3.25km, (f) SCREAM_1.625km, (g) SCREAM_H, and (h) WRF_RAP.

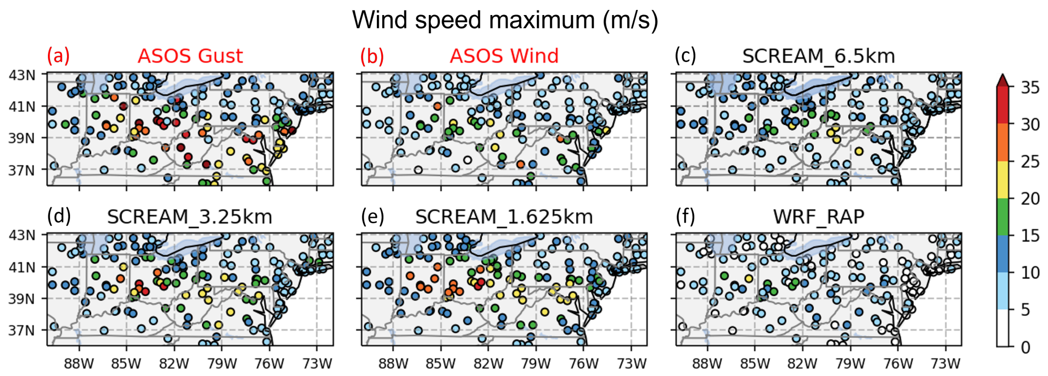

Figure 12. Wind speed maximum (m/s) between 18:00 UTC 29 June and 06:00 UTC 30 June 2012 in (a) ASOS gust, (b) ASOS wind, (c) SCREAM_6.5km, (d) SCREAM_3.25km, (e) SCREAM_1.625km, and (f) WRF_RAP. Only ASOS sites with gust reports during the period are displayed. Panels (c-f) show results at the closest gird points to the ASOS locations.

man correlations with 2-hour shift do not display significant differences among simulations (with difference < 0.03). The simulations have larger areas with higher accumulated precipitation amount (> 15 mm) than observations, illustrated by BS larger than 1, and consistent with Figure 9c. Comparing SCREAM_H to SCREAM_3.25km, the H simulation shows worse BS, indicating the area with greater accumulated precipitation amount (> 15 mm) is enhanced when employing the H dynamical core.

Figure 12 shows 10-m wind speed maximum between 18:00 UTC 29 June and 06:00 UTC 30 June 2012 in ASOS records and simulations. Only ASOS sites with gust reports during the period are displayed. Figure 13 shows the histogram of wind speed maximum to quantify the number of stations with wind speed maximum in each 5 m/s interval.

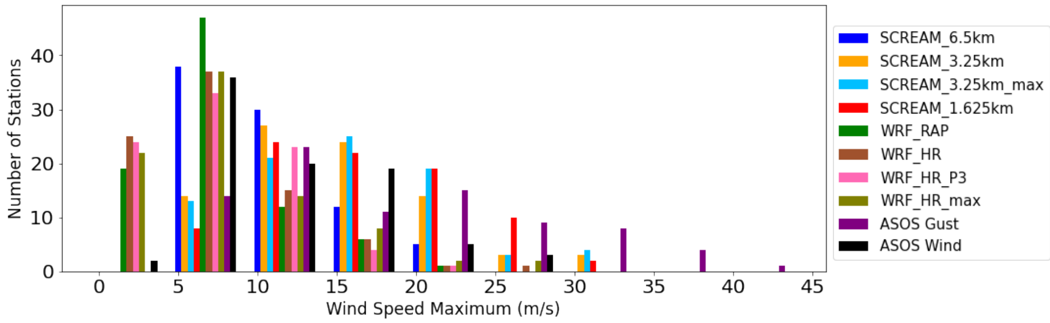

Figure 13. Histogram of wind speed maximum between 18:00 UTC 29 June and 06:00 UTC 30 June 2012 in SCREAM_6.5km (dark blue), SCREAM_3.25km (yellow), SCREAM_3.25km_max (light blue), SCREAM_1.625km (red), WRF_RAP (green), WRF_HR (brown), WRF_HR_P3 (pink), WRF_HR_max (olive), ASOS gust (purple), and ASOS wind (black) in the analysis region.

Figures S2-S3 are the same as Figures 12-13 but with the two-hour shift, producing similar results.

The wind speed maximum in SCREAM is generally between the values reported by ASOS gust and ASOS wind. SCREAM produces wind speeds exceeding 25 m/s at 3.25 km and 1.625 km resolutions during approximately the first half of the derecho life cycle, over central Indiana and Ohio, while the wind speed maximum is lower (20-25 m/s) during the second half of the derecho life cycle over northern West Virginia when the derecho size and intensity decline (Figures 8 and 9). SCREAM_6.5km displays an opposite pattern with higher wind speeds in the second half of the derecho path, associated the under-development of the system before 00:00 UTC at coarser resolution (Figure 6b). WRF_RAP shows lower wind speeds than both ASOS gust and ASOS wind speeds, with maximum wind speed lower than 25 m/s. This result is consistent with the WRF result in Shepherd et al. (2021) (see their Figure 9) where the wind speed maximum is under 25 m/s. Given that wind damage generally occurs above 25.7 m/s and is proportional to the cube of the wind speed, the underestimation of 10-m wind speeds may make it difficult to leverage the WRF simulations to estimate wind damage during such storms.

In Figure 13, the distribution of wind speed maximum in SCREAM_1.625km agrees best with the ASOS gust showing most of the stations in the range of 10-25 m/s as well as $\sim$ 15 stations with extremely high speed ($>$ 25 m/s). None of the SCREAM simulations display the wind maximum lower than 5 m/s. As the SCREAM grid spacing decreases, the histogram shifts towards being right-skewed, representing generally higher wind speeds. For WRF_RAP, wind speed maximum higher 25 m/s is not represented. Further, it has twice as many stations with wind speed maximum lower than 10 m/s as in ASOS wind, but only 35-70% of the ASOS stations fall into categories greater than 10 m/s.

4.4 Sensitivity Tests of Wind Speeds

To better understand the discrepancy between SCREAM and WRF simulations, in term of 10-m wind speed (Figures 5, 10, 12 and 13), we consider three factors that might affect the simulated wind speed and conduct additional sensitivity tests to provide more insights into the wind speed in diverse model configurations.

Table 8. Regional-averaged 10-m wind speed maximum (m/s) between 18:00 UTC 29 June to 06:00 UTC 30 June 2012 in the analysis region (76°-88°W, 36.5°-42°N). The reference simulation in the calculation of percentage is shown in the parenthesis of the last column.

Simulation Name	Regional-averaged 10-m Wind Speed Maximum (m/s)	Increase of Regional-averaged 10-m Wind Speed Maximum (%)
WRF_RAP	8.99	–
WRF_HR	9.36	4.12 (WRF_RAP)
WRF_HR_P3	9.46	1.07 (WRF_HR)
WRF_HR_max	10.28	9.83 (WRF_HR)
SCREAM_6.5km	10.71	–
SCREAM_3.25km	15.52	–
SCREAM_3.25km_max	16.47	6.12 (SCREAM_3.25km)
SCREAM_1.625km	16.62	–

Firstly, the SCREAM HR configuration (used in all SCREAM simulations except for SCREAM_LR; Table 1) uses 128 vertical levels (92 levels below 100 hPa), while WRF_RAP uses 45 vertical levels (all below 100 hPa), which may cause higher 10-m wind speed in SCREAM than WRF. A new WRF simulation run with 72 vertical levels (WRF_HR; Table 4) is performed and shown in Figure 14a. Comparing WRF_HR with Figure 12f, increasing the number of vertical levels does lead to higher 10-m wind speed, especially in central Indiana and southern Ohio. Table 8 quantifies the regional-averaged 10-m wind speed maximum. WRF_HR has a similar wind maximum pattern as WRF_RAP but with an increased regional-averaged wind maximum by 4.12%. The histogram of wind speed maximum in the new WRF_HR is also shown in Figure 13. More vertical levels reduce the wind speed maximum in the 5-10 m/s bin by 30% and slightly increases frequency of high wind speed (> 25 m/s). However, the WRF_HR wind speed maximum distribution still exhibits a low bias when compared with the ASOS wind in all categories above 10 m/s.

Secondly, the simulated 10-m wind speed is also related to the microphysical scheme applied since the microphysical scheme affects the convective structure and the cold pool associated with it. A new simulation WRF_HR_P3 (Figure 14b,Table 4) is conducted by replacing the microphysical scheme in WRF_HR with P3, which produces an insignificant increase in regionally-averaged wind maximum (1%). The P3 scheme's impact is more noticeable in shifting the derecho propagation path southward by 1-3° rather than modifying the 10-m wind speed magnitude exclusively.

Thirdly, both SCREAM and WRF results in the previous analyses are instantaneous outputs at 15-minute frequency, which possibly do not capture the highest wind speed during the 15-minute period. Therefore, we further output the wind speed maximum during each 15-minute period in SCREAM_3.25km and WRF_HR, labeled as SCREAM_3.25km_max (Figure 14d) and WRF_HR_max (Figure 14c), respectively. This change causes an increase of regional-averaged wind maximum by 9.83% in WRF and 6.12% in SCREAM (Table 8), becoming the most influential factor in this section to the wind speed. This suggests that future work involving the assessment of 10-m wind hindcast against high-frequency observations (such as 5-minute ASOS) should consider a higher output frequency of the wind speed than other variables. The highest wind speed in SCREAM_3.25km_max then exceeds 30 m/s. WRF_HR_max display the most stations with wind speed maximum above 25 m/s among all WRF simulations. Although the wind speed increases (by 9.83%) in WRF_HR_max than WRF_RAP, there is an underestimation of wind speed in every category above 10 m/s comparing WRF_HR_max to ASOS gust or wind (Figure 13). Specifically, the total number of stations with wind speeds > 10 m/s is 47 in ASOS wind, 71 in ASOS gust, and 26 in WRF_HR_max (less than ASOS wind and gust by 45% and 63%, respectively).

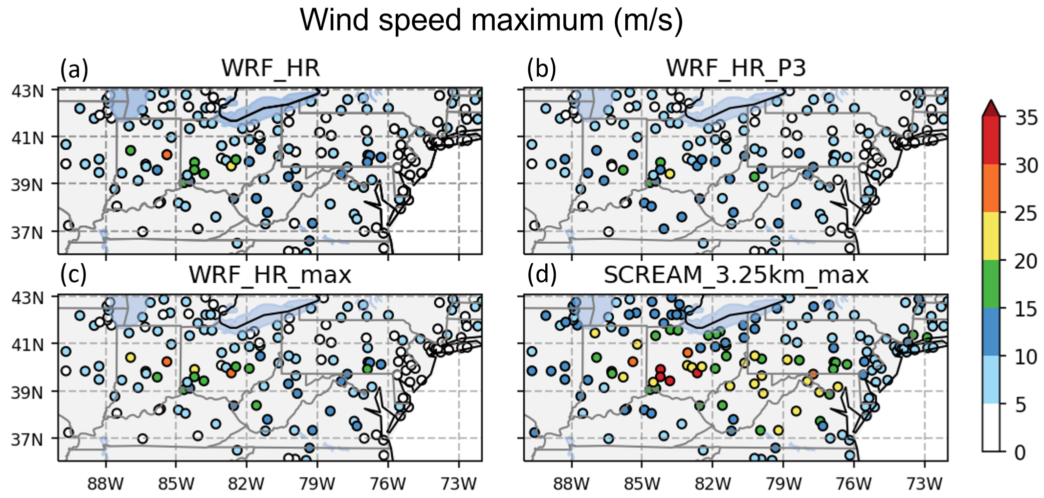

Figure 14. Same as Figure 12 but in (a) WRF_HR, (b) WRF_HR_P3, (c) WRF_HR_max, and (d) SCREAM_3.25km_max.

5 Conclusions

Climate models have been evaluated primarily using average behavior over large areas or long time periods (Gleckler et al., 2008; Eyring et al., 2019). However, it is rare for evaluations to consider the fidelity of simulating the most extreme and high-impact weather phenomena. Evaluating the capability of models to reproduce poorly predicted, but severe historic events will facilitate further model development and comparison, enable optimization of model configuration, and provide context for examining future changes in such events. In this work, we present one such extreme event testbed for evaluating climate modeling systems that operate at cloud resolving scales. The testbed focuses on the hindcast of the June 2012 North American derecho, which is one of the most devastating and strongly under-forecasted events in the US, with the intention of providing a largely comprehensive suite of diagnostics and statistical metrics for model assessment and intercomparison.

The metrics aim to assess the spatiotemporal characteristics of the derecho using RRM approach in SCREAM at various resolutions (Table 1) and regional WRF model at 4km (Table 4) against observations. Sensitivity tests address RRM grid spacing ranging from 6.5 - 1.625 km, differences between hydrostatic and nonhydrostatic dynamical cores, low-resolution and high-resolution model configurations, initialization time, and source for the ICs (i.e., RAP, ERA5, and ERAI). Besides OLR, precipitation, and composite radar reflectivity fields, this study places additional emphasis on the 10-m wind speed evaluation, which has not been thoroughly covered in previous studies. It is worth mentioning that the metrics package evaluated here is independent of the tracking method.

The representation of the derecho is shown to benefit from the finer horizontal resolutions in SCREAM, particularly at 1.625 km grid spacing, in a variety of ways including: cold cloud temperature and its coverage, radar reflectivity structure of the MCS, derecho position during the propagation, and simulated surface gust wind speed. The derecho-associated cold cloud in the SCREAM simulation at the coarsest resolution (6.5 km) is significantly underdeveloped with a smaller coverage maximum by $\sim 23\%$ and a longer lag (~ 2 hours) in the peak time compared to simulations at finer resolution. These results reinforce the need for higher resolution in operational convection-permitting models.

The simulations exhibit high dependence on the IC source and the initialization time, revealing the initial environment to be one of the most important factors for the simulation quality. Although it is impossible to determine the most superior product, as results may vary on a case by case basis, in this study RAP is found to be the best choice of IC source. Simulations initialized with RAP provide significantly improved performance in both SCREAM and WRF, compared with ERA5, ERAI, and NARR (Figures 2-4) for this specific event. In particular, SCREAM simulations with ERA5 and ERAI are not able to generate a realistic organized convection pattern.

The SCREAM HR model configuration (no deep convection scheme and more vertical model levels) produces a significantly better storm than the LR model configuration, which fails to develop an organized precipitating system over the affected region (Figures 2-4). The simulation with the hydrostatic dynamical core is similar to the non-hydrostatic one when examining individual snapshots (Figures 2-4) but shows greater biases ($< 6\%$) in the averaged OLR over the 12-hour period (Table 6).

While both SCREAM and WRF models show high pattern correlations (> 0.88) between the simulated OLR and the observation (Table 6), SCREAM is characterized by lower cloud top (indicated by 33-42% more biases in OLR RMSE; Table 6) and smaller cold cloud coverage by up to 50% than WRF (Figure 8), especially in the second half of the derecho life cycle.

SCREAM and WRF simulations both capture the observed derecho track, but both produce a delay of approximately 2 hours in feature location and associated gust front timing (Figures 6 and 10). Among all simulations, SCREAM at 1.625 km resolution displays the smallest time lag with an difference of $\sim$ 0.5-1 hour from the 3.25 km simulation. Both models overestimate the precipitation intensity over the precipitating grid points (up to 100% in WRF and 45% in SCREAM; Figure 9b) and the areas with composite radar reflectivity > 40 dBZ (up to 4 times in both models; Figure 8b), and underestimate the precipitating area ($\sim$ 70% in WRF and 47% in SCREAM; Figure 8c). WRF yields higher wet biases (up to 20% higher in accumulated precipitation RMSE; Table 7 and Figure 9c) but over smaller precipitation feature by $\sim$ 45% than SCREAM (Figure 8c). The overall bias magnitudes of 12-hour accumulated precipitation in the models fall in the range of CMORPH and IMERG compared to the NCEP Stage IV precipitation, except for a higher RMSE in WRF (Table 7). Our results highlight the importance of using multiple metrics to reveal different aspects of the simulations and errors.

SCREAM captures the bow-shape echo with a tilted axis more realistically than WRF (Figure 3). Moreover, the largest discrepancies between SCREAM and WRF are apparent in the 10-m wind speed. SCREAM simulates a 10-m wind speed maximum in between ASOS wind and ASOS gust speeds and a highest wind speed above 30 m/s, significantly higher than WRF by $\sim$ 73% (Table 8). WRF underestimates the wind speed maximum compared to either ASOS wind (by 27-70%) or gust speeds (by 56-85%; Figure 10) and does not produce damaging wind speeds > 25 m/s (Figure 13). Further investigation shows that this underestimation of the 10-m wind speed in WRF could be partly reduced by finer vertical resolution (4.12%) or changing the analyzed output from the 15-minute instantaneous model result to the maximum during each 15-minute interval (9.83%; Table 8).

Last but not least, we suggest some potential applications for future studies. SCREAM RRM demonstrates competitive utility for studying individual high-impact weather events when compared to a high-resolution regional climate model (WRF), and so could be employed for future regional climate model simulations. We argue that it could be useful for assessing and tuning resolution-dependent configurations in global models and for short-term weather prediction at fine scales (Zarzycki & Jablonowski, 2015). We further expect the extreme weather testbed described here is useful for future cloud-resolving model intercomparisons, such as to models from the DYnamics of the Atmospheric general circulation On Non-hydrostatic Domains (DYAMOND) project (Stevens et al., 2019), performed in similar hindcast mode. This suite of assessment will be useful in objectively evaluating model design choices related to extreme weather phenomenon, building credibility for extreme event attribution, and developing physical climate storylines to explore plausible changes of extreme events in the future.

Appendix A Derecho tracking with TempestExtremes

For feature tracking in the simulations and observations, we use TempestExtremes 2.2.1 (Ullrich & Zarzycki, 2017; Ullrich et al., 2021). The exact commands employed in this analysis are provided here for reference.

```
$TEMPESTEXTREMESDIR/DetectBlobs --in_data FLUT.nc --out DetectBlobs.FLUT.nc
--thresholdcmd "FLUT,<,163,0" --geofiltercmd "area,>=,5000km2"
--lonname lon --latname lat --regional
```

```
$TEMPESTEXTREMESDIR/StitchBlobs --in DetectBlobs.FLUT.nc
--out StitchBlobs.FLUT.nc --var "binary_tag" --outvar "id" --mintime "6h"
--min_overlap_prev 50 --regional --lonname lon --latname lat
```

Data Availability Statement

SCREAM is available online (E3SM Project, 2022). Simulation results (including SCREAM and WRF) and scripts used to plot figures could be archived at Zenodo (`https://doi.org/10.5281/zenodo.6617206`).

Acknowledgments
This research was supported as part of the Energy Exascale Earth System Model (E3SM) project "The Simple Cloud-Resolving E3SM Atmosphere Model (SCREAM)", funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

We wish to thank the editor and anonymous reviewers for their suggestions to improve this manuscript. The authors would also like to thank Ben Hillman for his instructions about building new RRM grids in SCREAM.

References

AghaKouchak, A., Behrangi, A., Sorooshian, S., Hsu, K., & Amitai, E. (2011). Evaluation of satellite-retrieved extreme precipitation rates across the central united states. *Journal of Geophysical Research: Atmospheres*, *116*(D2).

AghaKouchak, A., Mehran, A., Norouzi, H., & Behrangi, A. (2012). Systematic and random error components in satellite precipitation data sets. *Geophysical Research Letters*, *39*(9).

Alliss, R., & Hoffman, M. (2010). Quasi-linear convective system mesovorticies and tornadoes. *Meteorology Program, Iowa State University, Ames*.

Anthes, R. A. (1983). Regional models of the atmosphere in middle latitudes. *Monthly weather review*, *111*(6), 1306–1335.

Ban, N., Schmidli, J., & Schär, C. (2014). Evaluation of the convection-resolving regional climate modeling approach in decade-long simulations. *Journal of Geophysical Research: Atmospheres*, *119*(13), 7889–7907.

Beck, H. E., Pan, M., Roy, T., Weedon, G. P., Pappenberger, F., Van Dijk, A. I., ... Wood, E. F. (2019). Daily evaluation of 26 precipitation datasets using stage-iv gauge-radar data for the conus. *Hydrology and Earth System Sciences*, *23*(1), 207–224.

Beck, H. E., Vergopolan, N., Pan, M., Levizzani, V., Van Dijk, A. I., Weedon, G. P., . . . Wood, E. F. (2017). Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling. *Hydrology and Earth System Sciences*, *21*(12), 6201–6217.

Benjamin, S. G., Weygandt, S. S., Brown, J. M., Hu, M., Alexander, C. R., Smirnova, T. G., . . . others (2016). A north american hourly assimilation and model forecast cycle: The rapid refresh. *Monthly Weather Review*, *144*(4), 1669–1694.

Bernardet, L. R., & Cotton, W. R. (1998). Multiscale evolution of a derecho-producing mesoscale convective system. *Monthly weather review*, *126*(11), 2991–3015.

Bertagna, L., Guba, O., Taylor, M. A., Foucar, J. G., Larkin, J., Bradley, A. M., . . . Salinger, A. G. (2020). A Performance-Portable Nonhydrostatic Atmospheric Dycore for the Energy Exascale Earth System Model Running at Cloud-Resolving Resolutions. In *Sc20: International conference for high performance computing, networking, storage and analysis* (pp. 1–14).

Bowman, K. P., & Homeyer, C. R. (2017). *Gridrad - three-dimensional gridded nexrad wsr-88d radar data.* Boulder CO: Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. Retrieved from https://doi.org/10.5065/D6NK3CR7

Brockhaus, P., Luthi, D., & Schar, C. (2008). Aspects of the diurnal cycle in a regional climate model. *Meteorologische Zeitschrift*, *17*(4), 433–444.

Brousseau, P., Seity, Y., Ricard, D., & Léger, J. (2016). Improvement of the forecast of convective activity from the arome-france system. *Quarterly Journal of the Royal Meteorological Society*, *142*(699), 2231–2243.

Caldwell, P. M., Mametjanov, A., Tang, Q., Van Roekel, L. P., Golaz, J.-C., Lin, W., . . . others (2019). The DOE E3SM coupled model version 1: Description and results at high resolution. *Journal of Advances in Modeling Earth Systems*, *11*(12), 4095–4146.

Caldwell, P. M., Terai, C. R., Hillman, B., Keen, N. D., Bogenschutz, P., Lin, W., . . . others (2021). Convection-permitting simulations with the e3sm global atmosphere model. *Journal of Advances in Modeling Earth Systems*, *13*(11), e2021MS002544.

Carey, L. D., Murphy, M. J., McCormick, T. L., & Demetriades, N. W. (2005). Lightning location relative to storm structure in a leading-line, trailing-stratiform mesoscale convective system. *Journal of Geophysical Research: Atmospheres*, *110*(D3).

Chen, F., & Dudhia, J. (2001). Coupling an advanced land surface–hydrology model with the penn state–ncar mm5 modeling system. part i: Model implementation and sensitivity. *Monthly weather review*, *129*(4), 569–585.

Corfidi, S. F., Coniglio, M. C., Cohen, A. E., & Mead, C. M. (2016). A proposed revision to the definition of "derecho". *Bulletin of the American Meteorological Society*, *97*(6), 935–949.

Dai, A., Giorgi, F., & Trenberth, K. E. (1999). Observed and model-simulated diurnal cycles of precipitation over the contiguous united states. *Journal of Geophysical Research: Atmospheres*, *104*(D6), 6377–6402.

Davis, C., Brown, B., & Bullock, R. (2006). Object-based verification of precipitation forecasts. part i: Methodology and application to mesoscale rain areas. *Monthly Weather Review*, *134*(7), 1772–1784.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., . . . others (2011). The era-interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the royal meteorological society*, *137*(656), 553–597.

Demaria, E., Rodriguez, D., Ebert, E., Salio, P., Su, F., & Valdes, J. (2011). Evaluation of mesoscale convective systems in south america using multiple satellite products and an object-based approach. *Journal of Geophysical Research: Atmospheres*, *116*(D8).

Dennis, J., Fournier, A., Spotz, W. F., St-Cyr, A., Taylor, M. A., Thomas, S. J., & Tufo, H. (2005). High-resolution mesh convergence properties and parallel efficiency of a spectral element atmospheric dynamical core. *The International Journal of High Performance Computing Applications*, *19*(3), 225–235.

Dennis, J. M., Edwards, J., Evans, K. J., Guba, O., Lauritzen, P. H., Mirin, A. A., ... Worley, P. H. (2012). CAM-SE: A scalable spectral element dynamical core for the Community Atmosphere Model. *The International Journal of High Performance Computing Applications*, *26*(1), 74–89.

Du, J. (2011). *Ncep/emc u.s. stage iv imagery. version 1.0.* UCAR/NCAR - Earth Observing Laboratory. Retrieved from https://doi.org/10.26023/V0KW-KG0V-9J0W

Dye, J., Winn, W., Jones, J., & Breed, D. (1989). The electrification of new mexico thunderstorms: 1. relationship between precipitation development and the onset of electrification. *Journal of Geophysical Research: Atmospheres*, *94*(D6), 8643–8656.

E3SM Project. (2022). *Simple cloud-resolving e3sm atmosphere model (scream).* Retrieved from https://github.com/E3SM-Project/scream/releases/tag/SCREAM_Derecho_RRM_JAMES2022

Elliott, E. J., Yu, S., Kooperman, G. J., Morrison, H., Wang, M., & Pritchard, M. S. (2016). Sensitivity of summer ensembles of fledgling superparameterized us mesoscale convective systems to cloud resolving model microphysics and grid configuration. *Journal of Advances in Modeling Earth Systems*, *8*(2), 634–649.

Evans, K. J., Lauritzen, P., Mishra, S., Neale, R., Taylor, M., & Tribbia, J. (2013). Amip simulation with the cam4 spectral element dynamical core. *Journal of Climate*, *26*(3), 689–709.

Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G., Caldwell, P., ... others (2019). Taking climate model evaluation to the next level. *Nature Climate Change*, *9*(2), 102–110.

Feng, Z., Leung, L. R., Houze Jr, R. A., Hagos, S., Hardin, J., Yang, Q., ... Fan, J. (2018). Structure and evolution of mesoscale convective systems: Sensitivity to cloud microphysics in convection-permitting simulations over the united states. *Journal of Advances in Modeling Earth Systems*, *10*(7), 1470–1494.

Feng, Z., Leung, L. R., Liu, N., Wang, J., Houze Jr, R. A., Li, J., ... Guo, J. (2021). A global high-resolution mesoscale convective system database using satellite-derived cloud tops, surface precipitation, and tracking. *Journal of Geophysical Research: Atmospheres*, *126*(8), e2020JD034202.

Feng, Z., Song, F., Sakaguchi, K., & Leung, L. R. (2021). Evaluation of mesoscale convective systems in climate simulations: Methodological development and results from mpas-cam over the united states. *Journal of Climate*, *34*(7), 2611–2633.

Fierro, A. O., Gao, J., Ziegler, C. L., Mansell, E. R., MacGorman, D. R., & Dembek, S. R. (2014). Evaluation of a cloud-scale lightning data assimilation technique and a 3dvar method for the analysis and short-term forecast of the 29 june 2012 derecho event. *Monthly Weather Review*, *142*(1), 183–202.

Figurski, M., Nykiel, G., Jaczewski, A., Bałdysz, Z., & Wdowikowski, M. (2017). The impact of initial and boundary conditions on severe weather event simulations using a high-resolution wrf model. case study of the derecho event in poland on 11 august 2017. *Meteorology Hydrology and Water Management*.

Fiolleau, T., & Roca, R. (2013). An algorithm for the detection and tracking of tropical mesoscale convective systems using infrared images from geostationary satellite. *IEEE transactions on Geoscience and Remote Sensing*, *51*(7), 4302–4315.

Fosser, G., Khodayar, S., & Berg, P. (2015). Benefit of convection permitting climate model simulations in the representation of convective precipitation. *Climate Dynamics*, *44*(1), 45–60.

Fujita, T. T. (1978). Manual of downburst identification for project nimrod(atmospheric circulation). *Satellite and Mesometeorology Research Paper*, 104.

Gao, Y., Leung, L. R., Zhao, C., & Hagos, S. (2017). Sensitivity of us summer precipitation to model resolution and convective parameterizations across gray zone resolutions. *Journal of Geophysical Research: Atmospheres*, *122*(5), 2714–2733.

Gleckler, P. J., Taylor, K. E., & Doutriaux, C. (2008). Performance metrics for climate models. *Journal of Geophysical Research: Atmospheres*, *113*(D6).

Golaz, J.-C., Caldwell, P. M., Van Roekel, L. P., Petersen, M. R., Tang, Q., Wolfe, J. D., . . . others (2019). The DOE E3SM coupled model version 1: Overview and evaluation at standard resolution. *Journal of Advances in Modeling Earth Systems*, *11*(7), 2089–2129.

Grim, J. A., Rauber, R. M., McFarquhar, G. M., Jewett, B. F., & Jorgensen, D. P. (2009). Development and forcing of the rear inflow jet in a rapidly developing and decaying squall line during bamex. *Monthly weather review*, *137*(4), 1206–1229.

Grunzke, C. T., & Evans, C. (2017). Predictability and dynamics of warm-core mesoscale vortex formation with the 8 may 2009 "super derecho" event. *Monthly Weather Review*, *145*(3), 811–832.

Guastini, C. T., & Bosart, L. F. (2016). Analysis of a progressive derecho climatology and associated formation environments. *Monthly Weather Review*, *144*(4), 1363–1382.

Haberlie, A. M., & Ashley, W. S. (2019). A radar-based climatology of mesoscale convective systems in the united states. *Journal of Climate*, *32*(5), 1591–1606.

Halverson, J. B. (2014). A mighty wind: The derecho of june 29, 2012. *Weatherwise*, *67*(4), 24–31.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., . . . others (2018). *ERA5 hourly data on pressure levels from 1979 to present.* Copernicus Climate Change Service (C3S) Climate Data Store (CDS). Retrieved from `https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-pressure-levels?tab=overview`

Hong, Y., Hsu, K.-L., Sorooshian, S., & Gao, X. (2004). Precipitation estimation from remotely sensed imagery using an artificial neural network cloud classification system. *Journal of Applied Meteorology*, *43*(12), 1834–1853.

Houze Jr, R. A. (2004). Mesoscale convective systems. *Reviews of Geophysics*, *42*(4).

Hu, H., Leung, L. R., & Feng, Z. (2020a). Observed warm-season characteristics of mcs and non-mcs rainfall and their recent changes in the central united states. *Geophysical Research Letters*, *47*(6), e2019GL086783.

Hu, H., Leung, L. R., & Feng, Z. (2020b). Understanding the distinct impacts of mcs and non-mcs rainfall on the surface water balance in the central united states using a numerical water-tagging technique. *Journal of Hydrometeorology*, *21*(10), 2343–2357.

Hu, H., Leung, L. R., & Feng, Z. (2021). Early warm-season mesoscale convective systems dominate soil moisture–precipitation feedback for summer rainfall in central united states. *Proceedings of the National Academy of Sciences*, *118*(43).

Huang, X., Hu, C., Huang, X., Chu, Y., Tseng, Y.-h., Zhang, G. J., & Lin, Y. (2018). A long-term tropical mesoscale convective systems dataset based on a novel objective automatic tracking algorithm. *Climate dynamics*, *51*(7), 3145–3159.

Huang, X., & Ullrich, P. A. (2017). The changing character of twenty-first-century precipitation over the western united states in the variable-resolution cesm. *Journal of Climate*, *30*(18), 7555–7575.

Huffman, G. J., Bolvin, D. T., Braithwaite, D., Hsu, K., Joyce, R., Xie, P., & Yoo, S.-H. (2015). Nasa global precipitation measurement (gpm) integrated multi-satellite retrievals for gpm (imerg). *Algorithm Theoretical Basis Document (ATBD) Version*, *4*, 26. Retrieved from `https://disc.gsfc.nasa.gov/datasets/GPM_3IMERGDF_06/summary?keywords=%22IMERG%20final%22`

Iacono, M. J., Delamere, J. S., Mlawer, E. J., Shephard, M. W., Clough, S. A., & Collins, W. D. (2008). Radiative forcing by long-lived greenhouse gases: Calculations with the aer radiative transfer models. *Journal of Geophysical Research: Atmospheres*, *113*(D13).

Janjić, Z. I. (1994). The step-mountain eta coordinate model: Further developments of the convection, viscous sublayer, and turbulence closure schemes. *Monthly weather review*, *122*(5), 927–945.

Janowiak, J., Joyce, B., & Xie, P. (2017). *Ncep/cpc l3 half hourly 4km global (60s - 60n) merged ir v1 (gpm_mergir)*. Goddard Earth Sciences Data and Information Services Center (GES DISC). Retrieved from `https://disc.gsfc.nasa.gov/datasets/GPM_MERGIR_1/summary`

Johns, R. H., & Hirt, W. D. (1987). Derechos: Widespread convectively induced windstorms. *Weather and Forecasting*, *2*(1), 32–49.

Joyce, R. J., Janowiak, J. E., Arkin, P. A., & Xie, P. (2004). Cmorph: A method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution. *Journal of hydrometeorology*, *5*(3), 487–503.

Kharin, V. V., Zwiers, F. W., Zhang, X., & Hegerl, G. C. (2007). Changes in temperature and precipitation extremes in the ipcc ensemble of global coupled model simulations. *Journal of Climate*, *20*(8), 1419–1444.

Kolios, S., & Feidas, H. (2010). A warm season climatology of mesoscale convective systems in the mediterranean basin using satellite data. *Theoretical and applied climatology*, *102*(1), 29–42.

Lin, Y., & Mitchell, K. E. (2005). 1.2 the ncep stage ii/iv hourly precipitation analyses: Development and applications. In *Proceedings of the 19th conference hydrology, american meteorological society, san diego, ca, usa* (Vol. 10).

Liu, W., Ullrich, P. A., Guba, O., Caldwell, P. M., & Keen, N. D. (2022). An assessment of nonhydrostatic and hydrostatic dynamical cores at seasonal time scales in the energy exascale earth system model (e3sm). *Journal of Advances in Modeling Earth Systems*, *14*(2), e2021MS002805.

Maddox, R. A. (1980). Mesoscale convective complexes. *Bulletin of the American Meteorological Society*, 1374–1387.

Mesinger, F., DiMego, G., Kalnay, E., Mitchell, K., Shafran, P. C., Ebisuzaki, W., . . . others (2006). North american regional reanalysis. *Bulletin of the American Meteorological Society*, *87*(3), 343–360.

Miao, C., Ashouri, H., Hsu, K.-L., Sorooshian, S., & Duan, Q. (2015). Evaluation of the persiann-cdr daily rainfall estimates in capturing the behavior of extreme precipitation events over china. *Journal of Hydrometeorology*, *16*(3), 1387–1396.

Mittermaier, M., & Roberts, N. (2010). Intercomparison of spatial forecast verification methods: Identifying skillful spatial scales using the fractions skill score. *Weather and Forecasting*, *25*(1), 343–354.

Moker Jr, J. M., Castro, C. L., Arellano Jr, A. F., Serra, Y. L., & Adams, D. K. (2018). Convective-permitting hindcast simulations during the north american monsoon gps transect experiment 2013: Establishing baseline model performance without data assimilation. *Journal of Applied Meteorology and*

Climatology, *57*(8), 1683–1710.
Morrison, H., & Milbrandt, J. A. (2015). Parameterization of cloud microphysics based on the prediction of bulk ice particle properties. part i: Scheme description and idealized tests. *Journal of the Atmospheric Sciences*, *72*(1), 287–311.
Na, Y., Fu, Q., Leung, L. R., Kodama, C., & Lu, R. (2022). Mesoscale convective systems simulated by a high-resolution global nonhydrostatic model over the united states and china. *Journal of Geophysical Research: Atmospheres*, *127*(7), e2021JD035916.
Nadolski, V. (1998). Automated surface observing system (asos) user's guide. *National Oceanic and Atmospheric Administration, Department of Defense, Federal Aviation Administration, United States Navy*, *20*. Retrieved from `https://www.ncei.noaa.gov/pub/data/asos-fivemin/6401-2012/`
Nelson, B. R., Prat, O. P., Seo, D.-J., & Habib, E. (2016). Assessment and implications of ncep stage iv quantitative precipitation estimates for product intercomparisons. *Weather and Forecasting*, *31*(2), 371–394.
Peters, O., Neelin, J. D., & Nesbitt, S. W. (2009). Mesoscale convective systems and critical clusters. *Journal of the atmospheric sciences*, *66*(9), 2913–2924.
Pinto, J. O., Grim, J. A., & Steiner, M. (2015). Assessment of the high-resolution rapid refresh model's ability to predict mesoscale convective systems using object-based evaluation. *Weather and Forecasting*, *30*(4), 892–913.
Prein, A., Gobiet, A., Suklitsch, M., Truhetz, H., Awan, N., Keuler, K., & Georgievski, G. (2013). Added value of convection permitting seasonal simulations. *Climate Dynamics*, *41*(9), 2655–2677.
Prein, A. F., Langhans, W., Fosser, G., Ferrone, A., Ban, N., Goergen, K., … others (2015). A review on regional convection-permitting climate modeling: Demonstrations, prospects, and challenges. *Reviews of geophysics*, *53*(2), 323–361.
Prein, A. F., Liu, C., Ikeda, K., Bullock, R., Rasmussen, R. M., Holland, G. J., & Clark, M. (2020). Simulating north american mesoscale convective systems with a convection-permitting climate model. *Climate Dynamics*, *55*(1), 95–110.
Przybylinski, R. W. (1995). The bow echo: Observations, numerical simulations, and severe weather detection methods. *Weather and Forecasting*, *10*(2), 203–218.
Rhoades, A. M., Ullrich, P. A., & Zarzycki, C. M. (2018). Projecting 21st century snowpack trends in western usa mountains using variable-resolution cesm. *Climate Dynamics*, *50*(1), 261–288.
Roberts, N. (2008). Assessing the spatial and temporal variation in the skill of precipitation forecasts from an nwp model. *Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling*, *15*(1), 163–169.
Roberts, N. M., & Lean, H. W. (2008). Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, *136*(1), 78–97.
Sadeghi, M., Nguyen, P., Naeini, M. R., Hsu, K., Braithwaite, D., & Sorooshian, S. (2021). Persiann-ccs-cdr, a 3-hourly 0.04° global precipitation climate data record for heavy precipitation studies. *Scientific Data*, *8*(1), 1–11.
Sakaguchi, K., Leung, L. R., Zhao, C., Yang, Q., Lu, J., Hagos, S., … Lauritzen, P. H. (2015). Exploring a multiresolution approach using amip simulations. *Journal of Climate*, *28*(14), 5549–5574.
Sakaguchi, K., Lu, J., Leung, L. R., Zhao, C., Li, Y., & Hagos, S. (2016). Sources and pathways of the upscale effects on the southern hemisphere jet in mpas-cam4 variable-resolution simulations. *Journal of Advances in Modeling Earth Systems*, *8*(4), 1786–1805.
Salamanca, F., Krpo, A., Martilli, A., & Clappier, A. (2010). A new building energy model coupled with an urban canopy parameterization for urban climate simu-

lations—part i. formulation, verification, and sensitivity analysis of the model. *Theoretical and applied climatology*, *99*(3), 331–344.

Santer, B., Wigley, T., & Jones, P. (1993). Correlation methods in fingerprint detection studies. *Climate Dynamics*, *8*(6), 265–276.

Schenkman, A. D., & Xue, M. (2016). Bow-echo mesovortices: A review. *Atmospheric Research*, *170*, 1–13.

Schoen, J. M., & Ashley, W. S. (2011). A climatology of fatal convective wind events by storm type. *Weather and forecasting*, *26*(1), 109–121.

Schumacher, R. S., & Clark, A. J. (2014). Evaluation of ensemble configurations for the analysis and prediction of heavy-rain-producing mesoscale convective systems. *Monthly Weather Review*, *142*(11), 4108–4138.

Schumacher, R. S., & Johnson, R. H. (2005). Organization and environmental properties of extreme-rain-producing mesoscale convective systems. *Monthly weather review*, *133*(4), 961–976.

Schumacher, R. S., & Johnson, R. H. (2006). Characteristics of us extreme rain events during 1999–2003. *Weather and Forecasting*, *21*(1), 69–85.

Schumacher, R. S., & Rasmussen, K. L. (2020). The formation, character and changing nature of mesoscale convective systems. *Nature Reviews Earth & Environment*, *1*(6), 300–314.

Shepherd, T. J., Letson, F. L., Barthelmie, R. J., & Pryor, S. C. (2021). How well are hazards associated with derechos reproduced in regional climate simulations? *Natural Hazards and Earth System Sciences Discussions*, 1–42.

Shourd, K. N. (2017). *The multi-scale dynamics of the 29-30 june 2012" super derecho"* (Unpublished doctoral dissertation). University of Nevada, Reno.

Shourd, K. N., & Kaplan, M. L. (2021). The multiscale dynamics of the 29 june 2012 super derecho. *Climate*, *9*(11), 155.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Liu, Z., Berner, J., … others (2019). A description of the advanced research wrf model version 4. *National Center for Atmospheric Research: Boulder, CO, USA*, *145*, 145.

Squitieri, B. J., & Gallus, W. A. (2020). On the forecast sensitivity of mcs cold pools and related features to horizontal grid spacing in convection-allowing wrf simulations. *Weather and Forecasting*, *35*(2), 325–346.

Stensrud, D. J., Wicker, L. J., Xue, M., Dawson II, D. T., Yussouf, N., Wheatley, D. M., … others (2013). Progress and challenges with warn-on-forecast. *Atmospheric Research*, *123*, 2–16.

Stevens, B., Satoh, M., Auger, L., Biercamp, J., Bretherton, C. S., Chen, X., … others (2019). Dyamond: the dynamics of the atmospheric general circulation modeled on non-hydrostatic domains. *Progress in Earth and Planetary Science*, *6*(1), 1–17.

Stevenson, S. N., & Schumacher, R. S. (2014). A 10-year survey of extreme rainfall events in the central and eastern united states using gridded multisensor precipitation analyses. *Monthly Weather Review*, *142*(9), 3147–3162.

Tang, Q., Klein, S. A., Xie, S., Lin, W., Golaz, J.-C., Roesler, E. L., … others (2019). Regionally refined test bed in e3sm atmosphere model version 1 (eamv1) and applications for high-resolution modeling. *Geoscientific Model Development*, *12*(7), 2679–2706.

Tao, W.-K., & Chern, J.-D. (2017). The impact of simulated mesoscale convective systems on global precipitation: A multiscale modeling study. *Journal of Advances in Modeling Earth Systems*, *9*(2), 790–809.

Taylor, M. A., Guba, O., Steyer, A., Ullrich, P. A., Hall, D. M., & Eldrid, C. (2020). An energy consistent discretization of the nonhydrostatic equations in primitive variables. *Journal of Advances in Modeling Earth Systems*, *12*(1).

Thompson, G., Field, P. R., Rasmussen, R. M., & Hall, W. D. (2008). Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. part ii: Implementation of a new snow parameterization. *Monthly Weather Review*,

136(12), 5095–5115.
Toll, V., Männik, A., Luhamaa, A., & Rõõm, R. (2015). Hindcast experiments of the derecho in estonia on 08 august, 2010: Modelling derecho with nwp model harmonie. *Atmospheric Research*, *158*, 179–191.
Trenberth, K. E., Berry, J. C., & Buja, L. E. (1993). *Vertical interpolation and truncation of model-coordinate data.* Citeseer.
Ullrich, P. A. (2014). *Squadgen: Spherical quadrilateral grid generator.* Retrieved from `https://climate.ucdavis.edu/squadgen.php`
Ullrich, P. A., Devendran, D., & Johansen, H. (2016). Arbitrary-order conservative and consistent remapping and a theory of linear maps: Part II. *Monthly Weather Review*, *144*(4), 1529–1549.
Ullrich, P. A., & Taylor, M. A. (2015). Arbitrary-order conservative and consistent remapping and a theory of linear maps: Part I. *Monthly Weather Review*, *143*(6), 2419–2440.
Ullrich, P. A., & Zarzycki, C. M. (2017). TempestExtremes: A framework for scale-insensitive pointwise feature tracking on unstructured grids. *Geoscientific Model Development*, *10*(3), 1069–1090.
Ullrich, P. A., Zarzycki, C. M., McClenny, E. E., Pinheiro, M. C., Stansfield, A. M., & Reed, K. A. (2021). TempestExtremes v2.1: a community framework for feature detection, tracking and analysis in large datasets. *Geoscientific Model Development Discussions*, 1–37. Retrieved from `https://github.com/ClimateGlobalChange/tempestextremes`
Van Weverberg, K., Vogelmann, A., Lin, W., Luke, E., Cialella, A., Minnis, P., ... Jensen, M. (2013). The role of cloud microphysics parameterization in the simulation of mesoscale convective system clouds and precipitation in the tropical western pacific. *Journal of the Atmospheric Sciences*, *70*(4), 1104–1128.
Vié, B., Nuissier, O., & Ducrocq, V. (2011). Cloud-resolving ensemble simulations of mediterranean heavy precipitating events: Uncertainty on initial conditions and lateral boundary conditions. *Monthly Weather Review*, *139*(2), 403–423.
Weisman, M. L., & Rotunno, R. (2004). "a theory for strong long-lived squall lines" revisited. *Journal of the Atmospheric Sciences*, *61*(4), 361–382.
Weisman, M. L., Skamarock, W. C., & Klemp, J. B. (1997). The resolution dependence of explicitly modeled convective systems. *Monthly Weather Review*, *125*(4), 527–548.
Weyn, J. A., & Durran, D. R. (2017). The dependence of the predictability of mesoscale convective systems on the horizontal scale and amplitude of initial errors in idealized simulations. *Journal of the Atmospheric Sciences*, *74*(7), 2191–2210.
Wheatley, D. M., Yussouf, N., & Stensrud, D. J. (2014). Ensemble kalman filter analyses and forecasts of a severe mesoscale convective system using different choices of microphysics schemes. *Monthly Weather Review*, *142*(9), 3243–3263.
Wu, C., Liu, X., Lin, Z., Rhoades, A. M., Ullrich, P. A., Zarzycki, C. M., ... Rahimi-Esfarjani, S. R. (2017). Exploring a variable-resolution approach for simulating regional climate in the rocky mountain region using the vr-cesm. *Journal of Geophysical Research: Atmospheres*, *122*(20), 10–939.
Xie, P., Joyce, R., Wu, S., Yoo, S.-H., Yarosh, Y., Sun, F., & Lin, R. (2017). Reprocessed, bias-corrected cmorph global high-resolution precipitation estimates from 1998. *Journal of Hydrometeorology*, *18*(6), 1617–1641.
Xie, P., Joyce, R., Wu, S., Yoo, S.-H., Yarosh, Y., Sun, F., & Lin, R. (2019). *Noaa climate data record (cdr) of cpc morphing technique (cmorph) high resolution global precipitation estimates, version 1.* NOAA National Centers for Environmental Information. Retrieved from `https://doi.org/10.25921/w9va-q159`
Xu, Z., Rhoades, A. M., Johansen, H., Ullrich, P. A., & Collins, W. D. (2018). An intercomparison of gcm and rcm dynamical downscaling for characterizing

the hydroclimatology of california and nevada. *Journal of Hydrometeorology*, *19*(9), 1485–1506.

Yang, G.-Y., & Slingo, J. (2001). The diurnal cycle in the tropics. *Monthly Weather Review*, *129*(4), 784–801.

Yuan, J., & Houze, R. A. (2010). Global variability of mesoscale convective system anvil structure from a-train satellite data. *Journal of Climate*, *23*(21), 5864–5888.

Zarzycki, C. M., & Jablonowski, C. (2014). A multidecadal simulation of atlantic tropical cyclones using a variable-resolution global atmospheric general circulation model. *Journal of Advances in Modeling Earth Systems*, *6*(3), 805–828.

Zarzycki, C. M., & Jablonowski, C. (2015). Experimental tropical cyclone forecasts using a variable-resolution global model. *Monthly Weather Review*, *143*(10), 4012–4037.

Zhang, G. J., & McFarlane, N. A. (1995). Sensitivity of climate simulations to the parameterization of cumulus convection in the canadian climate centre general circulation model. *Atmosphere-ocean*, *33*(3), 407–446.

Zhang, S., Fu, H., Wu, L., Li, Y., Wang, H., Zeng, Y., ... others (2020). Optimizing high-resolution community earth system model on a heterogeneous many-core supercomputing platform. *Geoscientific Model Development*, *13*(10), 4809–4829.

Zhang, X., Anagnostou, E. N., & Schwartz, C. S. (2018). Nwp-based adjustment of imerg precipitation for flood-inducing complex terrain storms: Evaluation over conus. *Remote Sensing*, *10*(4), 642.

Zipser, E. J., & Lutz, K. R. (1994). The vertical profile of radar reflectivity of convective cells: A strong indicator of storm intensity and lightning probability? *Monthly Weather Review*, *122*(8), 1751–1759.

Supporting Information for "The June 2012 North American Derecho: A testbed for evaluating regional and global climate modeling systems at cloud-resolving scales"

W. Liu[1], P.A. Ullrich[1], J. Li [2], C. Zarzycki [3], P. M. Caldwell[4], L. R. Leung [2],

Y. Qian [2]

[1]Department of Land, Air, and Water Resources, University of California-Davis, Davis, CA, USA

[2]Pacific Northwest National Laboratory, Richland, WA, USA

[3]Department of Meteorology and Atmospheric Science, Pennsylvania State University, University Park, PA, USA

[4]Lawrence Livermore National Lab, Livermore, CA, USA

Contents of this file

Corresponding author: Weiran Liu, Department of Land, Air, and Water Resources, University of California-Davis, Davis, CA, USA. (wraliu@ucdavis.edu)

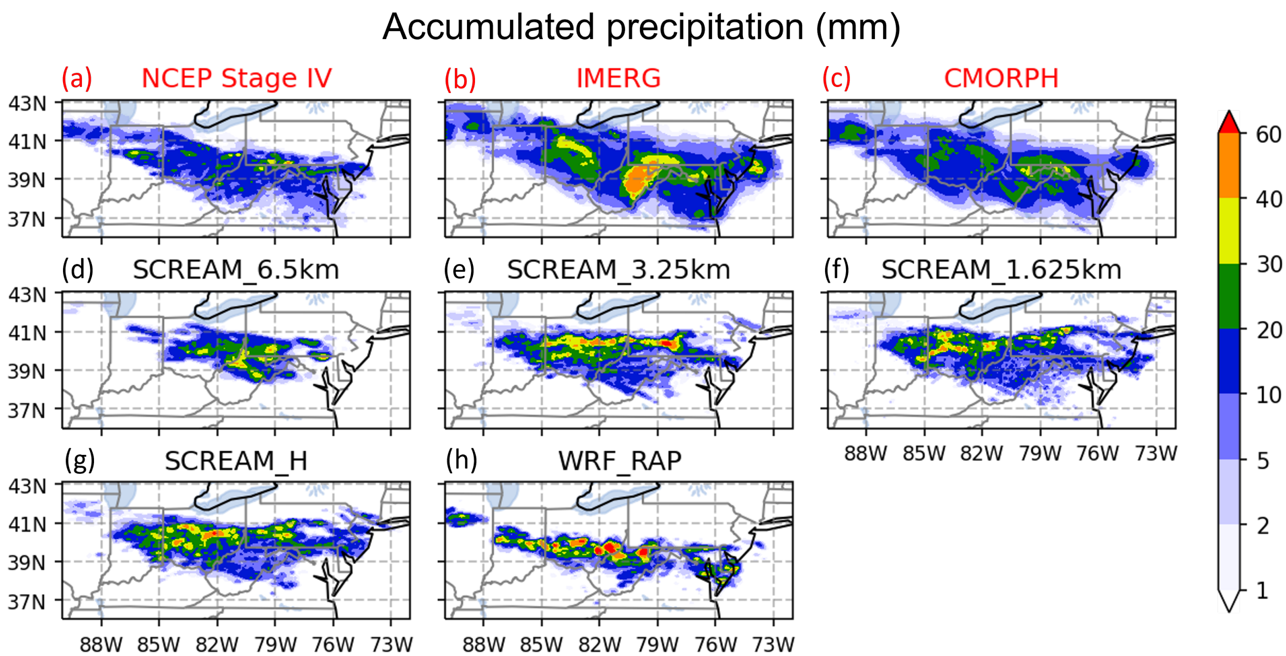

Figure S1. Same as Figure 11 but have a 2-hour shift for simulations resulting in the period changed to 20:00 UTC 29 June - 08:00 UTC 30 June 2012. Note that the time shift is applied only for simulation (i.e, panels a-e) and not applied for precipitation products (i.e., panels f-h).

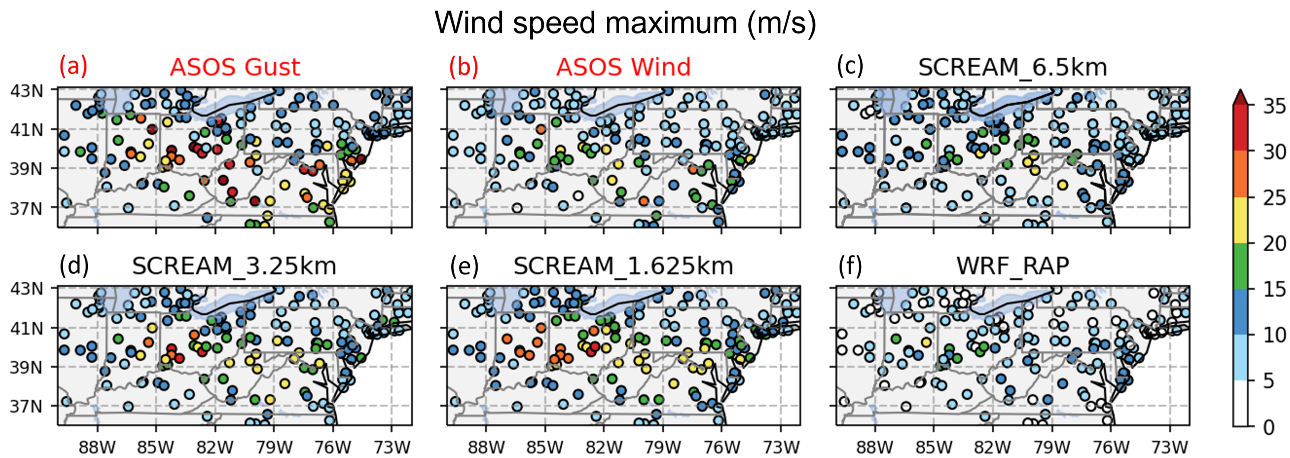

Figure S2. Same as Figure 12 but have a 2-hour shift for simulations resulting in the period changed to 20:00 UTC 29 June - 08:00 UTC 30 June 2012. Note that the time shift is applied only for simulation (i.e, panels a-e) and not applied for precipitation products (i.e., panels f-h).

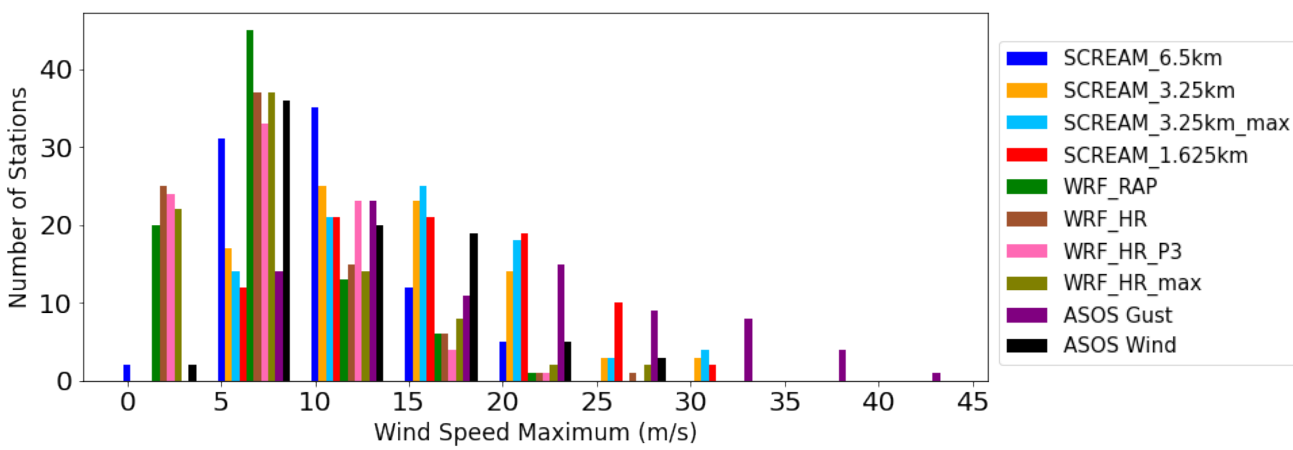

Figure S3. Same as Figure 13 but have a 2-hour shift for simulations resulting in the period changed to 20:00 UTC 29 June - 08:00 UTC 30 June 2012.

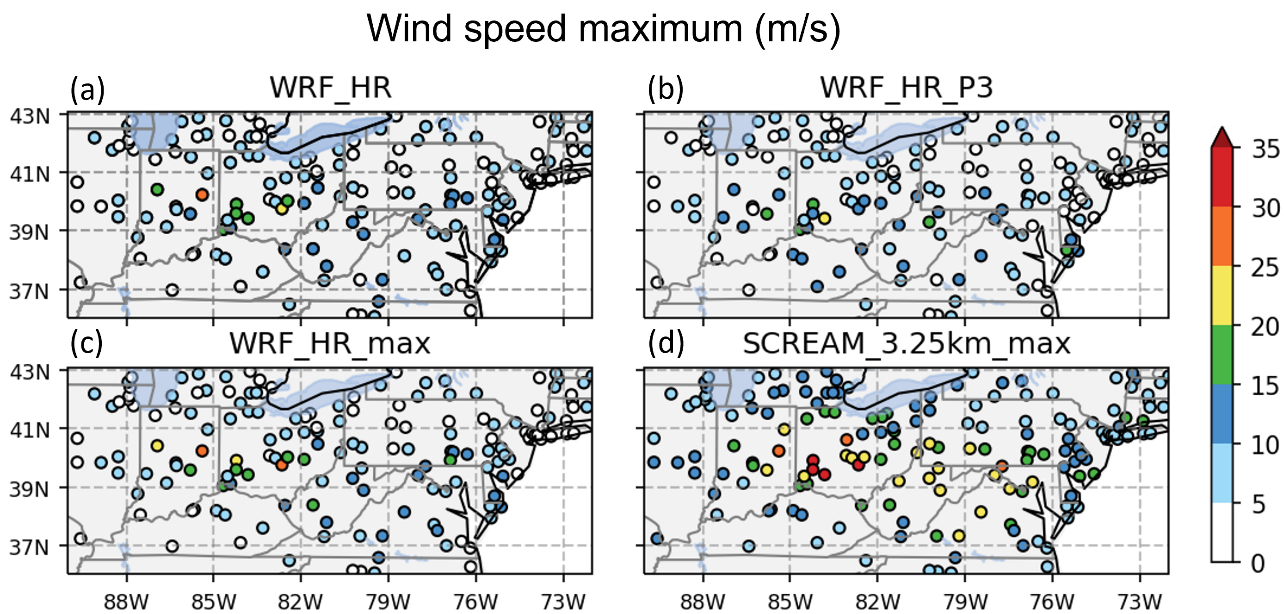

Figure S4. Same as Figure 14 but have a 2-hour shift for simulations resulting in the period changed to 20:00 UTC 29 June - 08:00 UTC 30 June 2012.