University of Reading

A comprehensive evaluation of biases in convective storm parameters in CMIP6 models over North America

Article

Accepted Version

Gopalakrishnan, D., Cuervo-Lopez, C., Allen, J. T., Trapp, R. J. and Robinson, E. (2025) A comprehensive evaluation of biases in convective storm parameters in CMIP6 models over North America. Journal of Climate, 38 (4). pp. 947-971. ISSN 1520-0442 doi: 10.1175/JCLI-D-24-0165.1 Available at https://centaur.reading.ac.uk/120543/

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

To link to this article DOI: http://dx.doi.org/10.1175/JCLI-D-24-0165.1

Publisher: American Meteorological Society

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

www.reading.ac.uk/centaur

University of Reading

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

LaTeX File (.tex, .sty, .cls, .bst, .bib)

Generated using the official AMS LaTeX template v6.1

A comprehensive evaluation of biases in convective-storm parameters in CMIP6 models over North America

Deepak Gopalakrishnan [a], Carlos Cuervo-Lopez [a], John T. Allen [a], Robert J. Trapp [b], and Eric Robinson [c]

[a] *Department of Earth and Atmospheric Sciences, Central Michigan University, Mount Pleasant, MI, USA,* [b] *Department of Atmospheric Sciences, University of Illinois at Urbana-Champaign, Urbana, IL, USA,* [c] *Aon Impact Forecasting, Chicago, IL, USA*

Corresponding author: Deepak Gopalakrishnan, dg.deepak@cmich.edu

Early Online Release: This preliminary version has been accepted for publication in *Journal of Climate*, may be fully cited, and has been assigned DOI 10.1175/JCLI-D-24-0165.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

© 2024 American Meteorological Society. This is an Author Accepted Manuscript distributed under the terms of the default AMS reuse license. For information regarding reuse and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

ABSTRACT: This study presents an evaluation of the skill of 12 global climate models from the CMIP6 archive in capturing convective-storm parameters over the United States. For the historical reference period 1979-2014, we compare the model-simulated 6-hourly CAPE, CIN, 0-1 km wind shear (S01) and 0-6 km wind shear (S06) to those from two independent reanalysis datasets – ERA5 and MERRA2. To obtain a comprehensive picture, we analyze the parameter distribution, climatological mean, extreme, and thresholded frequency of convective parameters. The analysis reveals significant bias in capturing both magnitude and spatial patterns, which also vary across the seasons. The spatial distribution of means and extremes of the parameters indicate that most models tend to overestimate CAPE, whereas S01, and S06 are underrepresented to varying extents. Additionally, models tend to underestimate extremes in CIN. Comparing the model profiles with rawinsonde profiles indicates that most of the high-CAPE models have warm and moist bias. We also find that the near-surface wind speed is generally underestimated by the models. The intermodel spread larger for thermodynamic parameters as compared to kinematic parameters. The models generally have a significant positive bias in CAPE over western and eastern regions of the continental US. More importantly, the bias in thresholded frequency of all four variables is considerably larger than the bias in mean, suggesting a non-uniform bias across the distribution. This likely leads to an under-representation of favorable severe thunderstorm environments, and has the potential to influence dynamical downscaling simulations via initial and boundary conditions.

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

SIGNIFICANCE STATEMENT: Global climate model projections are often used to explore future changes in severe thunderstorms activity. However, climate model outputs often have significant biases, and they can strongly impact the results. In this study, we thoroughly examined biases in convective parameters in 12 models from the Coupled Model Intercomparison Project Phase 6 with respect to two reanalysis datasets. The analysis is performed for North America, covering the period 1979-2014. The study reveals significant biases in convective parameters that differ between models, and are tied to the biases in temperature, humidity, and wind profiles. These results provide valuable insight into selecting the right set of models to analyze future changes in severe thunderstorm activity across the North American continent.

1. Introduction

The global mean surface temperature has increased approximately 1.1°C relative to pre-industrial baselines (1850-1900), primarily as a result of increased greenhouse gas emissions due to human activities (IPCC 2023). One of the direct consequence of tropospheric warming is generally an increase in water vapor content as dictated by the Clausius-Clapeyron (C-C) relationship, however studies have reported deviations from C-C scaling (e.g., O'Gorman and Muller 2010). This is expected to enhance thermodynamic instability, thereby increasing thunderstorm likelihood (Allen 2018; Trapp 2013). In the United States (US), severe convective storms (SCS) pose a substantial threat, resulting in significant loss of lives and billions of US dollars in insured losses annually (Doswell III 2003; Hoeppe 2016). Thus, how the risks associated with SCS might change in future assumes significant societal and scientific importance. However, while increases in boundary-layer moisture and low-level thermodynamic instability may favor the formation of SCS, modulating changes in convective inhibition and vertical wind shear may occur (Trapp et al. 2007; Brooks 2013; Diffenbaugh et al. 2013; Hoogewind et al. 2017; Lepore et al. 2021a; Haberlie et al. 2022). Due to this complex interplay, future changes in SCS have high uncertainty.

The bulk of the studies that investigate the future changes in SCS typically adopt two different approaches. Multiple global climate model (GCM) or regional climate model (RCM) are used to project how large-scale environments that promote formation and organization of SCS change in the future (Diffenbaugh et al. 2013; Allen et al. 2014a; Seeley and Romps 2015; Lepore et al. 2021a). The other approach involves performing climate-scale, convective-allowing numerical simulations

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

via "downscaling" the GCM outputs to examine the storm activity in future (Trapp et al. 2007, 2011; Gensini and Mote 2015; Hoogewind et al. 2017; Trapp et al. 2019; Rasmussen et al. 2020; Haberlie et al. 2022; Ashley et al. 2023). While GCMs use parameterization schemes to represent convection, convective-allowing simulations provide an explicit treatment of convection, thus adding greater value. Both the above approaches rely heavily on GCM outputs either directly or as initial/boundary conditions, and therefore the skill of the GCMs in capturing severe convective environments is of crucial importance. The analyses involving multiple model simulations generally reveal considerable inter-model spread in the simulated severe convective environments, though detailed analysis of these biases has generally not been explored (Trapp et al. 2007; Marsh et al. 2007; Seeley and Romps 2015; Lepore et al. 2021a; Chavas and Li 2022). Therefore, we argue that a more thorough evaluation of current generation GCMs is necessary before they are employed to explore changes in severe storm activity.

A majority of past evaluations have generally focused on justifying the use of GCMs for future projection of convective parameters, though rarely delve into the driving source of the biases beyond mean characteristics. Marsh et al. (2007) evaluated the performance of Community Climate System Model version 3 (CCSM3) in simulating severe thunderstorm environments over North America, comparing it to the NCEP/NCAR reanalysis dataset. Findings highlighted a significant underestimation of convective available potential energy (CAPE; hereinafter referred to as a thermodynamic parameter) over the continental US. Interestingly, the model-simulated CAPE values peaked over the Gulf of Mexico (GoM) and northern Atlantic, following the Gulf Stream, while demonstrating promising skill in reproducing severe thunderstorm activity over the central US. The analysis by Diffenbaugh et al. (2013) examining the Coupled Model Intercomparison Project phase 5 (CMIP5) models indicated significant improvement in terms of the representation of severe thunderstorm environments (defined based on the product of CAPE and 0-6km wind shear; CAPES06) over North America, however several models showed positive bias for CAPE in comparison to reanalysis data. Seeley and Romps (2015) also analyzed 11 CMIP5 models and noted that most models have excessive positive bias in CAPE, thereby impacting the magnitude and spatial distribution of severe thunderstorm environments. A recent study by Lepore et al. (2021a) compared the convective environments simulated by CMIP6 models and those derived from ERA-Interim reanalysis data. The multi-model mean consistently overestimated climatological CAPE

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

and underestimated 0-6 km wind shear climatology (S06; hereinafter referred to as a kinematic parameter) as compared to the reanalysis. Their study also noted that the climatological mean of convective inhibition (CIN; another thermodynamic parameter) is also slightly underestimated by the multi-model mean. Li et al. (2020a) evaluated the performance of the Community Atmosphere Model version 6 (CAM6) historical simulation (1980–2014) over North America. They observed that CAM6 simulations have a high CAPE bias in the eastern US as compared to ERA5, primarily resulting from near-surface moisture bias. They also noted that while climatological extreme patterns (99^{th} percentile) of S06 are well captured by CAM6, 0-3km storm-relative helicity is overestimated due to bias in low-level winds. More recently, Chavas and Li (2022) analyzed 13 models from the CMIP6 family and evaluated the models' skill in reproducing severe thunderstorm environments in comparison to ERA5 data in terms of CAPES06. The above study focused on higher percentiles of the CAPE and CAPES06 distributions. They noted that the bias in CAPES06 is primarily driven by bias in CAPE in models, and the study concluded that the bias CAPE is reflective of the bias in near-surface moist static energy in those models.

The studies investigating severe thunderstorm environments, in general, use the metric CAPES06 or a variant of it (e.g., Brooks et al. 2003; Trapp et al. 2007; Diffenbaugh et al. 2013; Seeley and Romps 2015; Chavas and Li 2022). Given that several GCMs tend to overestimate the CAPE fields, it is very likely that the bias, if any, in vertical wind shear field (S06) might be obscured or overlooked. Furthermore, the above-mentioned studies have not explored other convective parameters such as CIN and 0-1km wind shear (S01), a kinematic parameter, in detail nor biases in the variables that constitute these parameters. Considering that CIN and S01 are also key factors that modulate the formation and organization of severe thunderstorms (Brooks et al. 2003; Trapp 2013; Allen 2018), we argue that it is imperative to evaluate the models' skill in simulating these parameters, not only in terms of mean values but also in frequential space and their root causes. Therefore, the present study aims at providing a comprehensive evaluation of CMIP6 models in simulating the convective parameters such as CAPE, CIN, S01, and S06 over North America with respect to reanalysis data. To obtain a comprehensive picture, we analyze the total parameter distribution, mean, extreme, and thresholded frequency of convective parameters in CMIP6 models. We also analyze how bias varies at different percentiles of parameter distribution, which will show if the bias is skewed at higher or lower percentiles. Unlike earlier studies, which

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

primarily rely on a single reanalysis product, which may have its own inherit biases, here we compare the results from CMIP6 GCMs with two independent reanalysis products – ERA5 and the Modern-Era Retrospective Analysis for Research and Applications Version 2 (MERRA2). To the best of our knowledge, such an in-depth analysis of CMIP6 models featuring multiple aspects in the context of SCS does not exist in literature. Additionally, to better understand the biases in the driving output variables in each model, we compare the model-simulated temperature, humidity and wind fields with respect to those derived from rawinsonde soundings across North America. These results inform the choice of models for studying future SCS activity over North America either through convective environments, or using these data as initial conditions for downscaling.

2. Data and methods

a. CMIP6 and reanalysis datasets

The historical simulations from 12 GCMs from the CMIP6 archive (Eyring et al. 2016) are examined (Table 1). We selected models to cover a broad range of horizontal and vertical grid spacing and the selection was reliant on the availability of 3D profiles necessary to calculate 6-hourly convective variables. The ensemble used in the present study is slightly different from the one in Chavas and Li (2022). In addition to those analyzed in Chavas and Li (2022), the present study considers BCC-CSM2-MR, CESM2, FGOALS-g3, and NorESM2-MM. Whereas, AWI-ESM-1-1-LR and GFDL-CM4, included in Chavas and Li (2022), have not been utilized in our analysis owing to issues of availability for future projections. The convective parameterization scheme used in CMIP6 models is indicated in Table ST1.

The majority of the model outputs are obtained through the Google Cloud Public Data set. The datasets for CMCC-CM2-SR5 and NorESM2-MM were sourced through the Earth System Grid Federation (ESGF) data portal. We used the GCM outputs at 6-hourly temporal resolution covering the period 1979-2014.

Reanalysis products serve as common tools for assessing model performance, however they also exhibit biases, especially in convective environments (Allen et al. 2014b; Gensini and Mote 2014; King and Kennedy 2019; Li et al. 2020a; Wang et al. 2021; Taszarek et al. 2021b). When aiming to study long-term patterns in convective parameters, it is desirable to use multiple reference datasets, particularly in the light of known biases in many reanalysis products (e.g. Taszarek et al. 2021c;

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

TABLE 1. List of the CMIP6 models used in the study. Numbers in parenthesis in the No. of gridpoints entry are the number of vertical levels within the lowest 500 hPa for regions over the ocean.

Model name	Variant ID	No. of gridpoints [X × Y × Z]	References
BCC-CSM2-MR	r1i1p1f1	320 × 160 × 46 (11)	Wu et al. (2019)
CanESM5	r1i1p2f1	128 × 64 × 49 (18)	Swart et al. (2019)
CESM2	r11i1p1f1	288 × 192 × 32 (12)	Danabasoglu et al. (2020)
CMCC-CM2-SR5	r1i1p1f1	288 × 192 × 30 (12)	Cherchi et al. (2019)
CNRM-CM6	r11i1p1f2	256 × 128 × 91 (27)	Voldoire et al. (2019)
CNRM-ESM2	r11i1p1f2	256 × 128 × 91 (27)	Séférian et al. (2019)
FGOALS-g3	r3i1p1f1	180 × 80 × 26 (8)	Li et al. (2020b)
GISS-E2.1-G	r11i1p1f2	144 × 90 × 40 (14)	Kelley et al. (2020)
MIROC6	r1i1p1f1	256 × 128 × 81 (17)	Tatebe et al. (2019)
MPI-ESM1.2-HR	r1i1p1f1	384 × 192 × 95 (14)	Mauritsen et al. (2019)
MRI-ESM2.0	r1i1p1f1	320 × 160 × 80 (22)	Yukimoto et al. (2019)
NorESM2-MM	r1i1p1f1	288 × 192 × 32 (12)	Seland et al. (2020)

Pilguj et al. 2022). Therefore, in this study, we choose two independent reanalysis datasets – ERA5 and MERRA2 – for comparison with model outputs. This approach allows us to validate model-derived convective parameters against two widely accepted reanalysis datasets, enhancing the robustness and reliability of our findings.

ECMWF's ERA5 (Hersbach et al. 2020) is one of the popularly-used reanalysis products for model evaluation. ERA5 provides global atmospheric fields at a horizontal grid spacing of 0.25° × 0.25° on 137 terrain-following hybrid model levels (41 are within lowest 500mb) at 1-hourly temporal resolution. However, for consistency with the frequency of the CMIP6 models selected in this study, we utilized ERA5 outputs at 6-hourly frequency. MERRA2 (Gelaro et al. 2017) is another commonly used global atmospheric reanalysis product developed by the Global Modeling and Assimilation Office (GMAO), NASA. Compared to ERA5, MERRA2 comes on a coarser mesh with a horizontal grid spacing of 0.5° × 0.625° having 72 terrain-following hybrid vertical levels (22 are within lowest 500mb) and hence is more similar to the resolution of the GCMs. MERRA2 data is available from the year 1980, and we also used 6-hourly data to align with the temporal frequency of the GCMs.

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

b. Rawinsonde data and quality-checks

To provide an observational basis and supplement the environmental evaluation, rawinsonde profiles from the Integrated Global Radiosonde Archive version 2 (IGRA2) are used. IGRA2 is a collection of quality-controlled historical and near real-time rawinsonde and pilot balloon observations across the globe (Durre et al. 2018). Fig. SF1 shows the spatial distribution of sounding stations considered in this study. A second-level quality-check is performed for the rawinsonde observations following the methodology described in Taszarek et al. (2021c). This includes removal of records with missing variables, elimination of levels with temperature gradients higher than 10°C km^{-1}, and exclusion of values outside 1.5 times the interquartile range of the distribution. Rawinsonde profiles from 237 stations are considered here. Out of a total of 6377930 profiles, we used only 3476618 (54.51%) profiles that passed the quality check for the analysis here.

c. Convective parameters

The environments that are conducive for the formation of severe thunderstorms are typically inferred using parameters that summarize the atmospheric profile including CAPE, CIN, and vertical wind shear (Rasmussen and Blanchard 1998; Brooks et al. 2003; Trapp 2013). CAPE is a measure of instability of the atmosphere, which is computed as the total energy due to positive buoyancy of an air parcel between the level of free convection and the level of neutral buoyancy (Emanuel et al. 1994). CIN, on the other hand, corresponds to the negative buoyancy experienced by the air parcel and hence it essentially amounts to the energy required for an air parcel to reach LFC. Deep-layer wind shear (S06), computed as the magnitude of the vector difference in horizontal wind at 6 km height and the surface, is another crucial factor influencing the development of severe thunderstorms. The low-level wind shear (S01; magnitude of the vector difference in horizontal wind at 1 km and the surface) is a parameter that can impact the likelihood of tornadoes. Moreover, it allows us to assess reliability for near surface winds, a property that is known to be biased in earlier studies (Diffenbaugh et al. 2013; Allen et al. 2014b).

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

d. Methodology

We computed the above-mentioned convective parameters from CMIP6 models and reanalysis datasets for the period 1979-2014 (1980-2014 for MERRA2) at 6-hourly intervals on their native horizontal resolutions. CAPE and CIN are computed using the lowest 100 mb mixed-layer (ML) parcel for a pseudo-adiabatic ascent and applying the correction for virtual temperature following Doswell III and Rasmussen (1994). To compute S01 and S06, we first estimated the height above ground level using the hypsometric equation. Then, we interpolated wind vector to surface, 1- and 6-km above ground level and calculated the vector difference corresponding to S01 and S06. We used the python package *xcape* (Lepore et al. 2021b) to compute CAPE and CIN as it facilitates efficient calculation of the convective parameters by taking advantage of wrapped Fortran routines together with parallelization utilizing *dask*.

We compared the distributions of vertical profiles of temperature, specific humidity, and zonal and meridional wind components from GCMs with rawinsonde profiles. The observed and the model profiles are interpolated to 100 common height levels from surface to 40 km in vertical (above ground level), with 35 levels in the lower troposphere (0-2 km), for a consistent comparison between the varying model and observational level spacings. The distributions of observed soundings corresponding to 00Z and 12Z are compared with those from the GCMs at the nearest grid point in space and time. Additionally, profiles from the reanalysis fields are also compared to provide a reference baseline for the performance of reanalysis in capturing ambient meteorological features. The specific humidity profiles for MERRA2 are computed using temperature and relative humidity fields following Huang (2018).

3. Results

a. Parameter distributions

We analyze the distribution of convective parameters in terms of probability density functions (PDF) to assess the ensemble spread relative to the reanalysis fields. Analyzing the PDF helps gain insights into the general behavior of each ensemble member and the variability within the model simulations.

The model-simulated ML-CAPE exhibits considerable variability among the ensemble, and the spread is quite evident at the higher tail of the distribution (Fig. 1a). Among the CAPE

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

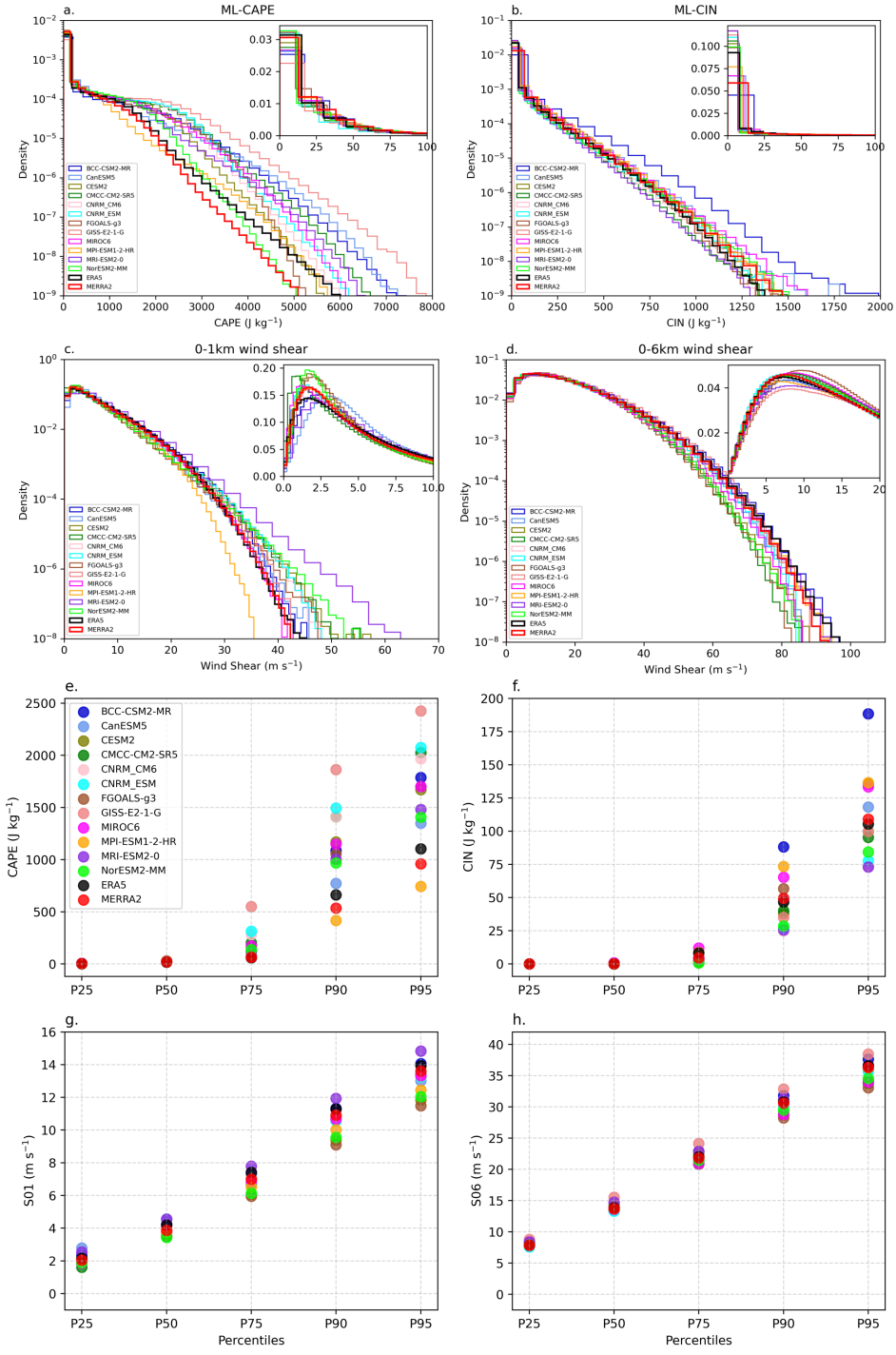

FIG. 1. PDF of (a) 100mb ML-CAPE, (b) 100mb ML-CIN, (c) S01, and (d) S06 from all grids points over the study region [130W-60W and 15N-60N] for the period 1980-2014 from 12 CMIP6 models, ERA5 and MERRA2. Datapoints are on their native grids at 6-hourly temporal resolution. The figures in inset show the same parameters for a shorter range to highlight the spread at the lower end of the PDF. Values corresponding to different percentiles from the above distribution for (e) ML-CAPE, (f) ML-CIN, (g) S01, and (h) S06 from all models are also shown.

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

distributions, MERRA2 is on the lowest end for CAPE values above 2000 J kg^{-1}, though MPI-ESM1-2-HR also shows significant underestimation. However, the majority of the models show excessive frequencies of high CAPE values, with GISS-E2.1-G, and CMCC-CM2-SR5 being on the higher end. Fig. 1e further highlights the inter-model spread at different CAPE percentiles. P95 CAPE from models spans from ~720 J kg^{-1} to ~2500 J kg^{-1}. GISS-E2.1-G shows high CAPE values for all higher percentiles. The summary statistics given in Table 2 indicate that MPI-ESM1-2-HR is the only model that underestimates the mean CAPE (~23% underestimation relative to ERA5), however it shows the least deviation from reanalysis CAPE. Other models have biases ranging from 39% (CanESM5) to 182% (GISS-E-2.1-G). We express the inter-model spread in terms of relative range of each parameter, which is defined as the ratio of ensemble range (the difference between maximum and minimum values among the ensemble) to the ensemble mean of respective parameters derived from 12 GCMs. The relative range for CAPE based on the mean values given in Table 2 is 1.185, meaning the ensemble range is larger than the ensemble mean. The standard deviation (SD) and upper-quartile difference (UQD; difference between 75$^{\text{th}}$ and 50$^{\text{th}}$ percentiles) in CAPE is comparatively large compared to reanalysis in most models (Table 2). UQD in models ranges from 44.27 J kg^{-1} in MPI-ESM1-2-HR to 515.11 J kg^{-1} in GISS-E-2.1-G, while that of ERA5 and MERRA2 is 41.97 J kg^{-1} and 44.04 J kg^{-1}, respectively. SD in CAPE ranges from 309 J kg^{-1} in MPI-ESM1-2-HR to 840 J kg^{-1} in GISS-E-2.1-G. UQD in MPI-ESM1-2-HR shows close agreement with the reanalysis, though SD is underestimated by 73 J kg^{-1} with respect to ERA5 (43 J kg^{-1} with respect to MERRA2). The analysis of CAPE values from Fig. 1 and Table 2 suggests that the model resolution does not always dictate performance for CAPE, for example, the higher-resolution BCC-CSM2-MR (~1.125° grid spacings) and a coarse-resolution model (CanESM5; ~2.8° grid spacings) have similar CAPE distribution.

ML-CIN distributions (Fig. 1b) show a smaller spread at higher values. An exception is BCC-CSM2-MR, which has higher CIN values almost throughout the distribution. The quantile values show (Fig. 1f) mixed bias for models as compared to the reanalysis-derived CIN values, with BCC-CSM2-SR5 having significantly larger values for the higher quantiles. While the mean and SD of CIN values from the models (Table 2) generally aligns with reanalysis, majority of the models exhibit a significant difference in UQD (ranges from 0.68 in NorESM2-MM to 11.12 in MIROC6). The difference between ERA5 and MERRA2 in terms of mean and SD of CIN is comparatively

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

TABLE 2. Mean, standard deviation (SD), and upper-quartile difference (UQD; difference between 75th and 50th percentile) of ML-CAPE, ML-CIN, S01, and S06 from 12 models and the two reanalysis products. The statistics are computed considering all grid points covering the study domain at a 6h temporal resolution on native grids during 1980-2014.

	ML-CAPE			ML-CIN			S01			S06		
Reanalysis/Model	Mean	SD	UQD	Mean	SD	UQD	Mean	SD	UQD	Mean	SD	UQD
ERA5	167.65	383.94	41.97	17.85	54.75	8.16	5.35	4.25	3.21	16.02	10.59	8.13
MERRA2	151.29	342.53	44.04	17.87	55.87	4.39	5.10	4.16	3.13	15.91	10.49	8.11
BCC-CSM2-MR	271.88	625.51	50.24	28.19	83.35	4.09	5.45	4.27	3.14	16.61	10.85	8.43
CanESM5	234.53	538.61	113.62	18.99	60.61	3.71	5.52	3.79	2.77	16.33	10.46	8.16
CESM2	284.23	562.52	147.95	15.51	54.35	0.85	4.55	3.67	2.59	15.53	9.83	7.59
CMCC-CM2-SR5	338.57	682.33	179.94	14.99	49.31	1.94	4.78	4.26	3.16	15.30	9.57	7.35
CNRM-CM6	340.28	657.83	258.90	12.99	47.31	2.66	4.96	4.13	3.03	15.67	10.40	7.98
CNRM-ESM	363.14	693.47	294.89	13.62	49.15	2.76	4.94	4.15	3.03	15.49	10.33	7.87
FGOALS-g3	270.65	590.66	96.74	19.82	62.12	3.28	4.49	3.53	2.43	15.44	9.26	7.01
GISS-E-2.1-G	472.28	840.11	515.11	15.87	55.45	2.86	5.13	4.10	3.11	17.51	11.07	8.59
MIROC6	288.39	592.41	145.67	22.64	62.05	11.12	4.93	4.13	3.08	15.42	9.68	7.31
MPI-ESM1.2-HR	128.70	309.53	44.27	23.13	64.21	8.66	4.77	3.77	2.78	16.23	10.51	8.16
MRI-ESM2.0	246.21	518.95	99.19	12.61	44.54	3.35	5.78	4.49	3.24	16.62	10.52	8.07
NorESM2-MM	239.57	477.74	113.63	13.58	51.65	0.68	4.54	3.74	2.69	15.62	9.89	7.72

small, but MERRA2 has higher values towards the tail of the distribution, which is reflected in lower UQD as compared to ERA5. Relative range for CIN based on model mean values (Table 2) is 0.882, which is smaller than that of CAPE.

The model S01 distributions are fairly consistent with the reanalyses in terms of the statistics in Table 2, though there are differences in the magnitude at which model-simulated S01 frequency peaks (Fig. 1c). The frequency peaks for CMCC-CM2-SR5 at $\sim$1.25 m s^{-1} (on lowest side) and for CanESM5, the frequency peaks at $\sim$3 m s^{-1} (on highest side). While most models overemphasize S01 magnitudes beyond 30 m s^{-1}, mean S01 values are slightly lower than that of the reanalyses. S01 values for various percentiles (Fig. 1g) display mixed biases in models as opposed to reanalysis, with MRI-ESM2-0 and FGOALS-g3 on the higher and lower side, respectively for higher percentiles. The inter-model spread in terms of relative range is 0.258, which is significantly lower than the relative range for both CAPE and CIN. The distributions of S06 (Table 2 and Fig. 1d) are also in agreement with the reanalysis datasets in terms of mean, SD, and UQD. However, majority of the models tend to underestimate the wind shear of values above 30 m s^{-1}. Fig. 1g suggests that

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

there is a greater consistency among the models, even at higher percentiles, though models such as MIROC6, FGOALS-g3, CMCC-CM2-SR5, NorESM2-MM display notable underestimation. Relative range for S06 is 0.138, which is the least among all convective variables. It is worthwhile to note that while S01 in MPI-ESM1.2-HR is underestimated, the S06 distribution is quite consistent with reanalyses.

We also examined the parameter distributions for all four seasons (Fig. SF2 – SF5). Though the general pattern of distribution remains similar to that of the annual distribution, the model bias does exhibit seasonal dependence.

b. Spatial pattern of extremes

To assess how the models capture the spatial patterns of the extremes, we examine the 95$^{\text{th}}$ percentile (P95) of all four parameters. Considering the higher frequency of severe thunderstorms over the US in spring and summer, and the apparent biases in these seasons, we focus primarily on these two seasons.

1) Spring (March–May)

While most models capture the general pattern of P95 ML-CAPE relative to the reanalysis fields (Fig. 2), there is noticeable bias in the magnitude, which is reflective of the positive bias we noted in the previous section (Table 2). The intermodel spread in terms of relative range for springtime P95 CAPE is 1.32, indicating that the range is larger than the mean. FGOALS-g3 shows lower CAPE values over the continental US. For instance, in the southeastern Texas, CAPE in ERA5 is around 2200 J kg^{-1}, whereas FGOALS-g3 indicates only around 1500 J kg^{-1}. As noted from the PDF (Fig.1a), MERRA2 shows lower values relative to ERA5, despite the location of the maxima is quite consistent between the reanalyses. CMCC-CM2-SR5, CNRM-CM6, CNRM-ESM2, and GISS-E2.1-G suffer from high bias over the GoM and southeast US. Similarly, CanESM5 shows higher values on the east coast (near the Carolinas). This type of bias is not unexpected in climate models given similar features in prior CMIP generations (e.g. Marsh et al. 2007; Diffenbaugh et al. 2013). Interestingly, CMCC-CM2-SR5 and MIROC6 exhibit a distinctive pattern characterized by a northwestward extension in higher CAPE values, reaching into the northern Great Plains. All models exhibit a pattern correlation higher than 0.8 with respect to both the reanalyses (Fig. 2o).

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

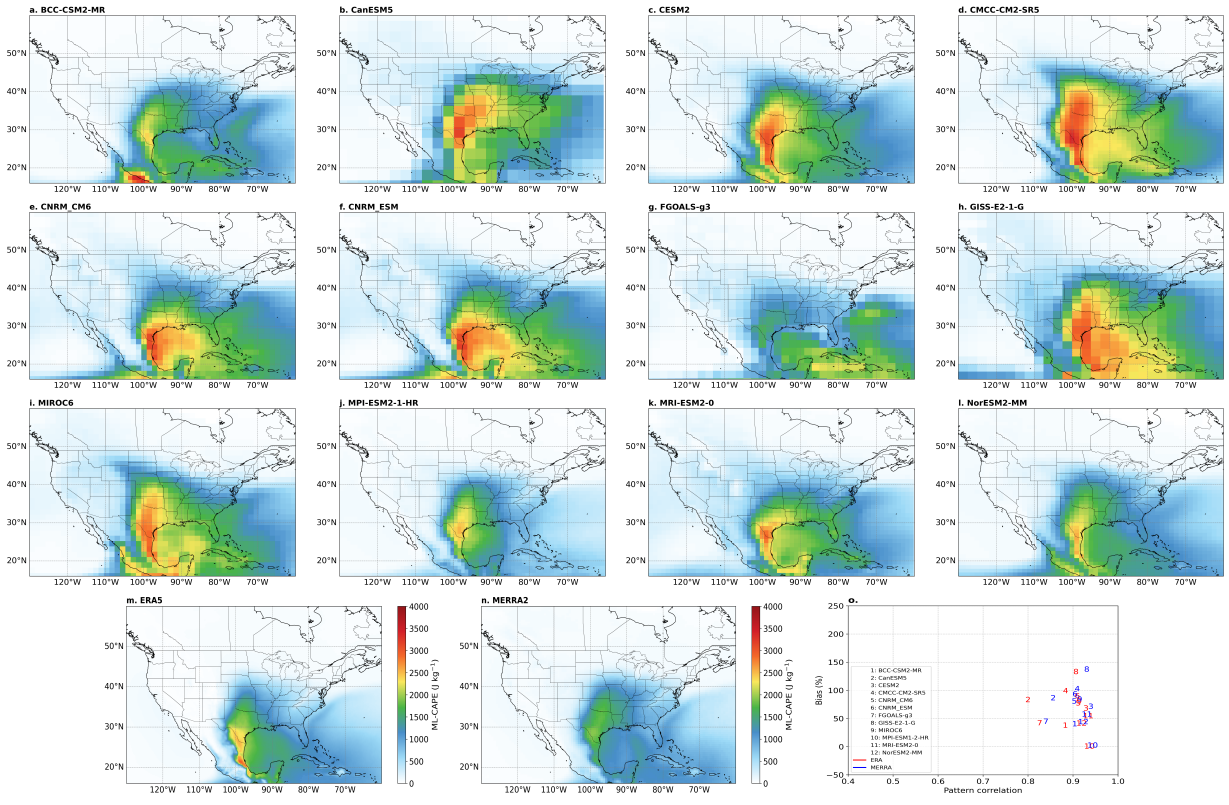

FIG. 2. Spatial map of 6-hourly P95 ML-CAPE for the spring season from (a-l) 12 CMIP6 models, along with (m) ERA5 and (n) MERRA2 reanalysis products covering the historical baseline period (1979-2014). The (o) scatter plot shows pattern correlation and bias for all 12 models with respect to (in red) ERA5 and (in blue) MERRA2. The bias and pattern correlation are computed after regridding all datasets to a uniform 1×1 lat-lon grid.

All other models show bias ~50% or higher. GISS-E2.1-G has a bias close to 150%, which is the largest among the ensemble. Better performance of MPI-ESM1-2-HR and excessive bias in GISS-E2.1-G during spring is in agreement with (Chavas and Li 2022). Intriguingly, models tend to correlate better with the spatial pattern of MERRA2 than ERA5, notably for CanESM5, CMCC-CM2-SR5, GISS-E2.1-G. MPI-ESM1-2-HR shows a near-zero bias.

P95 of ML-CIN (Fig. 3) has a pattern with maximum values concentrated over the western parts of the GoM, with an extension into the southern Great Plains. The P95 ML-CIN structure in both the reanalysis fields is quite similar, except that MERRA2 exhibits slightly higher values, primarily in the west side of the Rockies. While models are consistent at simulating the relative maxima over

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

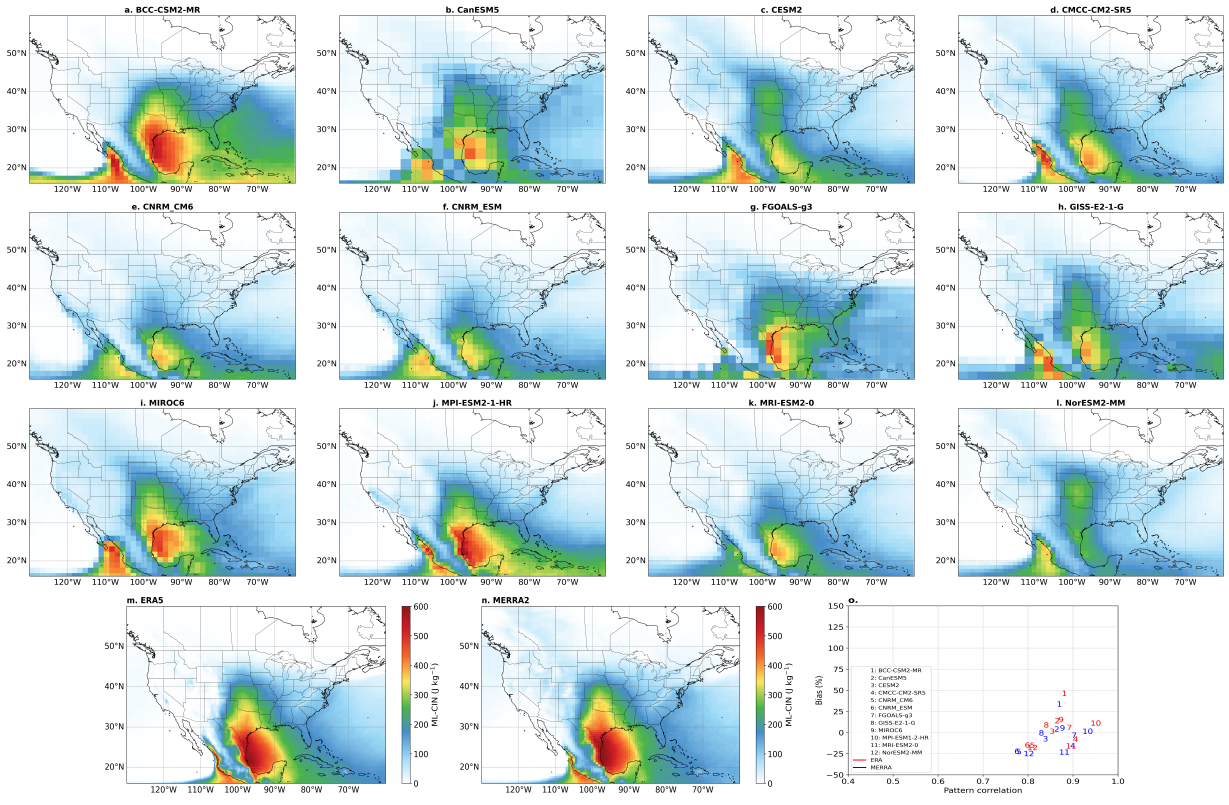

FIG. 3. As for Fig. 2 but for ML-CIN.

GoM and the southern Great Plains, most of them struggle to capture the magnitude accurately. The CNRM models and MRI-EMS2.0 fail to reproduce the higher values over the southern Great Plains. BCC-CSM2-MR exhibits the largest positive bias (close to 50%), whereas the two CNRM models, MRI-ESM2-0, and NorESM2-MM show approximately 25% underestimation in overall magnitude (Fig. 3o). MPI-ESM1.2-HR, again displays very good skill in capturing P95 CIN with a pattern correlation of >0.9 and bias close to 0, while MIROC6 shows better skill in simulating the P95 ML-CIN pattern relative to the reanalysis. The relative range for P95 CIN is 0.87, smaller than that of CAPE, suggesting that the intermodel spread is considerably smaller for P95 CIN as compared to that of CAPE.

P95 S01 has greater consistency between the models and reanalyses (Fig. 4) with smaller ensemble spread (relative range 0.29). As compared with ERA5, MERRA2 has slightly lower S01 values, particularly over the Great Plains and Midwest. Moreover, ERA5 exhibits higher shear values over high terrains of the Rockies owing to its finer horizontal grid spacing. Almost all models capture the spatial pattern well, as evidenced by a high pattern correlation (>0.9). We note

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

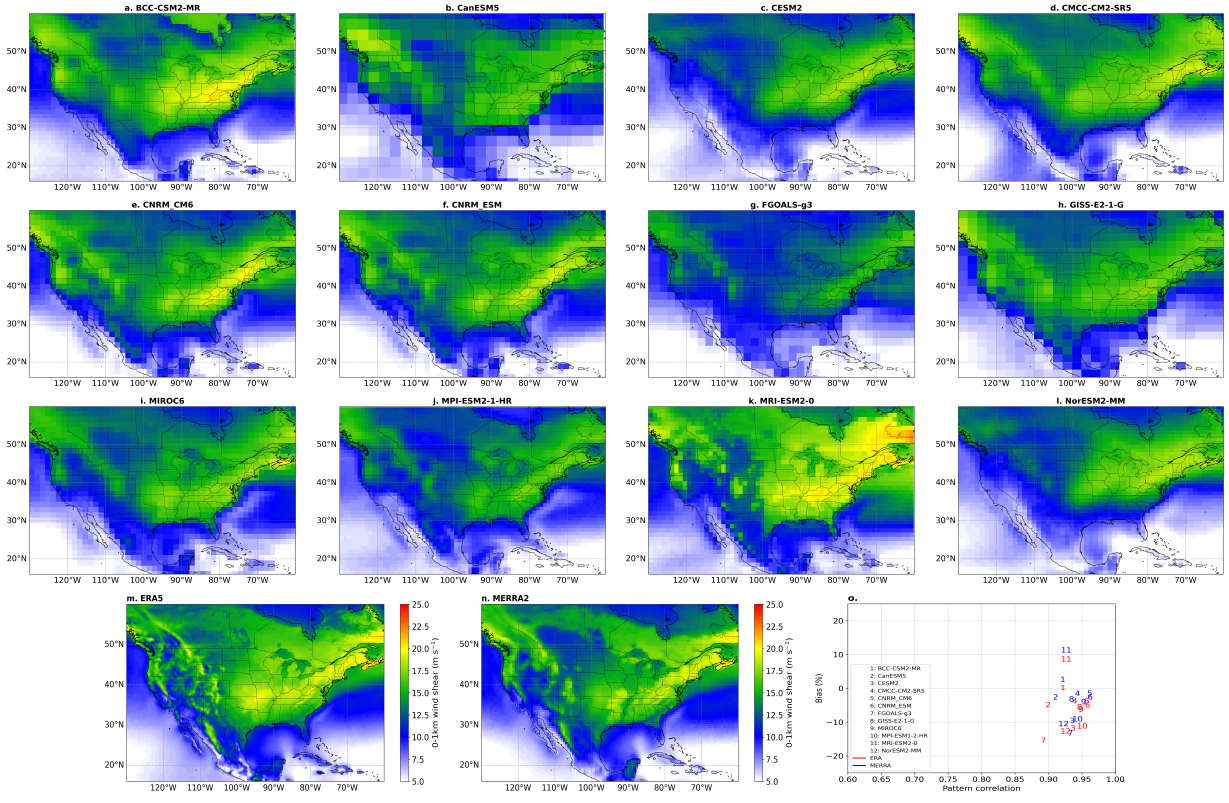

FIG. 4. As for Fig. 2 but for S01.

a slightly higher pattern correlation for most models relative to MERRA2 as compared to ERA5, likely due to the coarser horizontal grid spacing in MERRA2. MRI-ESM2.0 tends to overestimate the magnitude by approximately 10%. BCC-CSM2-MR, CanESM5, CMCC-CM2-SR5, CNRM models, and GISS-E2-1G show better skill capturing the S01 magnitude with bias less than 5%. All other models tend to underestimate P95 S01 with a bias close to 10%. FGOALS-g3 and MPI-ESM1.2-HR considerably underestimate P95 S01. Owing to similarities in the modeling components, both CESM2 and NorESM2-MM exhibit very similar pattern, and both these models do not capture the high shear values near the Rockies. Intriguingly, the model's skill in accurately capturing P95 S01 magnitude is not always dependent on the model grid spacing. For instance, CanESM5 has a much coarser mesh compared to CESM2, with approximate grid spacings of 2.8° and 1.25° degrees, respectively. However, CanESM5 exhibits a noticeably better skill in reproducing P95 S01 magnitude and spatial distribution (>0.9 pattern correlation and <5% bias).

The springtime P95 S06 distributions (Fig. 5) in the reanalysis products show consistent patterns, although ERA5 indicates slightly higher magnitudes compared to the values in MERRA2. The

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

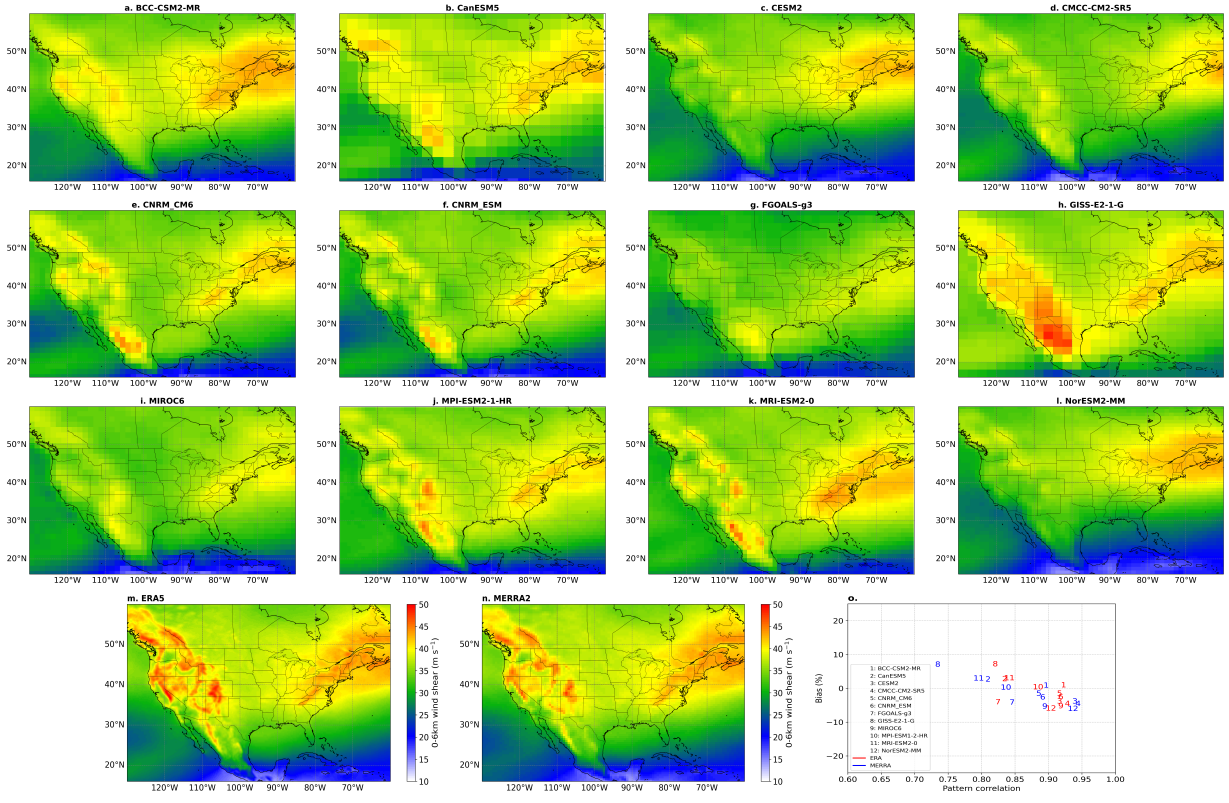

FIG. 5. As for Fig. 2 but for S06.

P95 S06 distribution is characterized by a pattern of higher S06 values along the Rockies and the west coast, with peaks reaching up to ~50 m s^{-1} in ERA5. S06 along the Appalachian Mountains and the northwest Atlantic is also stronger. While most models capture the pattern over Appalachian Mountains, the higher values near the Rockeis are underestimated. GISS-E2.1-G exhibits anomalously high S06 values over northern Mexico (up to 15 m s^{-1} higher values). CESM2, FGOALS-g3 and NorESM2-MM exhibit lower S06 magnitudes along the Rockies and west coast. While the bias across the ensemble lies within ±10%, the pattern correlation varies between 0.73 to 0.95. Curiously, the pattern correlation relative to ERA5 and MERRA2 is notably different in models, specifically for GISS-E2.1-G (0.83 and 0.73 respectively). The intermodel spread in terms of relative range for P95 S06 is 0.17, which is the least among the four variables during the spring season.

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

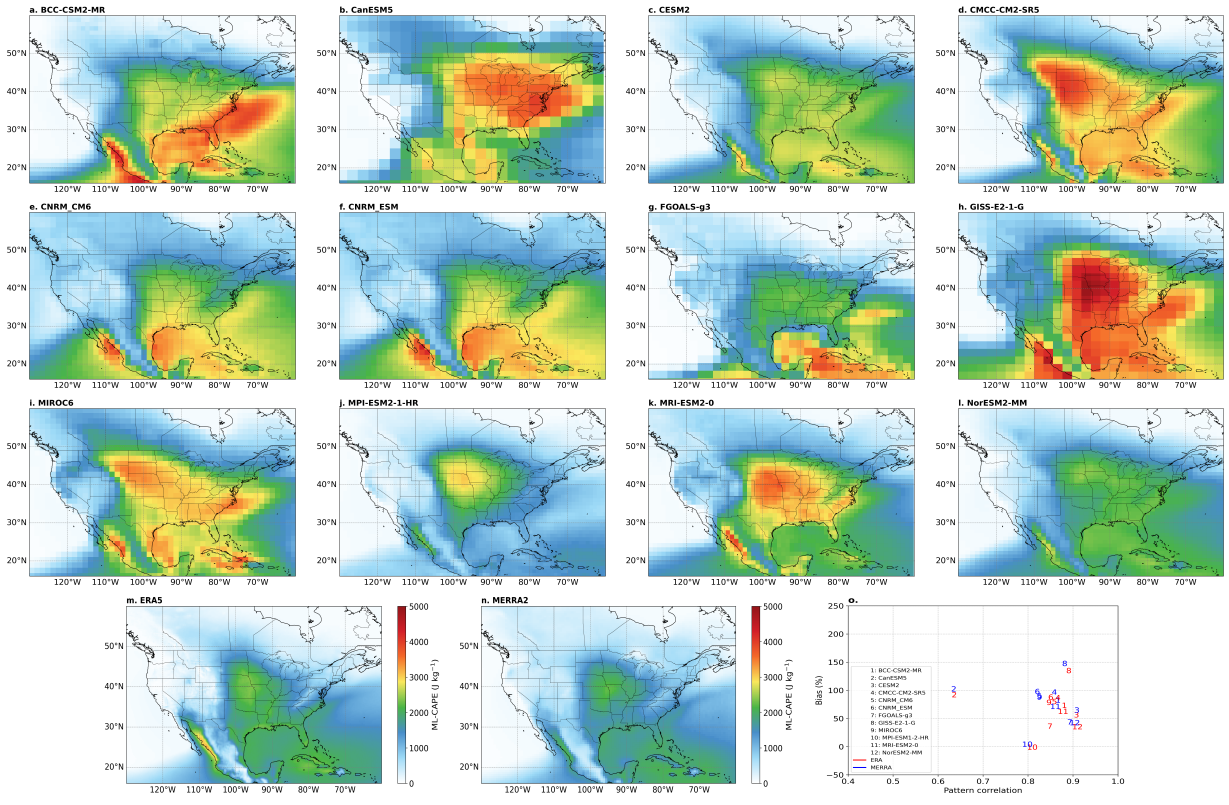

FIG. 6. Spatial map of 6-hourly P95 ML-CAPE for the summer season from 12 CMIP6 models, along with ERA5 and MERRA2 reanalysis products covering the historical baseline period (1979-2014).

2) SUMMER (JUNE–AUGUST)

In summer, most models heavily overestimate the P95 ML-CAPE values (Fig. 6) compared to reanalysis, similar to the findings in earlier work (Chavas and Li 2022). Among the reanalyses, MERAA2 shows slightly lower values overall and particularly near the Gulf of California (~1000 J kg^{-1} difference). While both of the reanalyses maximize ML-CAPE around 3000 J kg^{-1} near the Great Plains, several models have P95 ML-CAPE values higher than 4500 J kg^{-1}. The intermodel spread in terms of relative range of magnitude is 1.41, and this is higher as compared to that in spring (1.32). FGOALS-g3, MPI-ESM1.2-HR and NorESM2-MM are relatively better for CAPE magnitude (bias close to 50%), whereas GISS-E2.1-G is high outlier (150% bias). CMCC-CM2-SR5, CanESM5, GISS-E2.1-G, MIROC6, and MRI-ESM2.0 exhibit high CAPE in the Great Plains. As for spring, CMCC-CM2-SR5 and MIROC6 indicate the northwestward extension of CAPE maxima into the northern Great Plains in summer as well. Note that MPI-ESM1-2-HR,

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

unlike other models, have notably lower CAPE over the oceans. Intriguingly, CanESM5 displays a slightly different spatial pattern, with the maxima extending into the northeast US, resulting in poor pattern correlation (0.62). All other models have a pattern correlation higher than 0.8. The consistent high positive bias in models over the ocean, particularly along the northeast US coast is more pronounced in the summer compared to spring, suggesting biases in convective instability over warm waters. In contrast, the CCSM3 simulations in Marsh et al. (2007) displayed stronger bias near the northeast coast in spring.

The summertime P95 ML-CIN pattern (Fig. 7) has maxima over the Great Plains and near the Gulf of California. Compared to ERA5, MERRA2 exhibits slightly higher values over the Great Plains. Most models capture the P95 ML-CIN pattern similar to the reanalyses (pattern correlation >0.8), however the CNRM models and FGOALS-g3 are outliers. BCC-CSM2-MR shows much higher values, up to 800 J kg^{-1} (bias close to 100%), both the CNRM models exhibit markedly lower values over the Great Plains ($\sim$ 200 J kg^{-1}). FGOALS-g3 is another outlier showing much different spatial pattern as compared to other models and reanalyses, with maximum CIN values near the southern US. The relative range based on CIN magnitude from models is 1.3, indicating a large ensemble spread, which is notably larger than springtime CIN (0.87).

The spatial pattern of P95 S01 in summer (Fig. 8) is similar to spring. High S01 values are seen over the northeast and central US. MERRA2 exhibits slightly higher S01 values in the northern and northeastern regions compared to ERA5. As in spring, MRI-ESM2 slightly (<10% bias) overestimates the P95 S01 magnitude over the northeast US, though it captures the overall spatial structure reasonably well (0.92 pattern correlation). BCC-CSM2-MR, CESM2, MIROC6, and NorESM2 show relatively better skill in S01 spatial pattern and magnitude (> 0.85 pattern correlation and < 10% bias; Fig. 8o). However, CNRM-CM6, CNRM-ESM2, FGOALS-g3, and MPI-ESM1.2-HR show noticeably lower values relative to reanalysis (> 15% underestimation). This underestimation by the models is greater in summer relative to spring. The relative range for S01 magnitude is 0.27, which suggests smaller intermodel spread similar to that in spring.

In terms of P95 S06 (Fig. 9), the patterns in ERA5 and MERRA2 are almost identical with higher values, above 35 m s^{-1}, in the northwest region of the US. Higher S06 values are observed over the northeast US as well. BCC-CSM2-SR is an outlier which significantly overestimates S06 values in the northeast US and southern Canada. GISS-E-2.1-G and NorESM2-MM are the other two

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

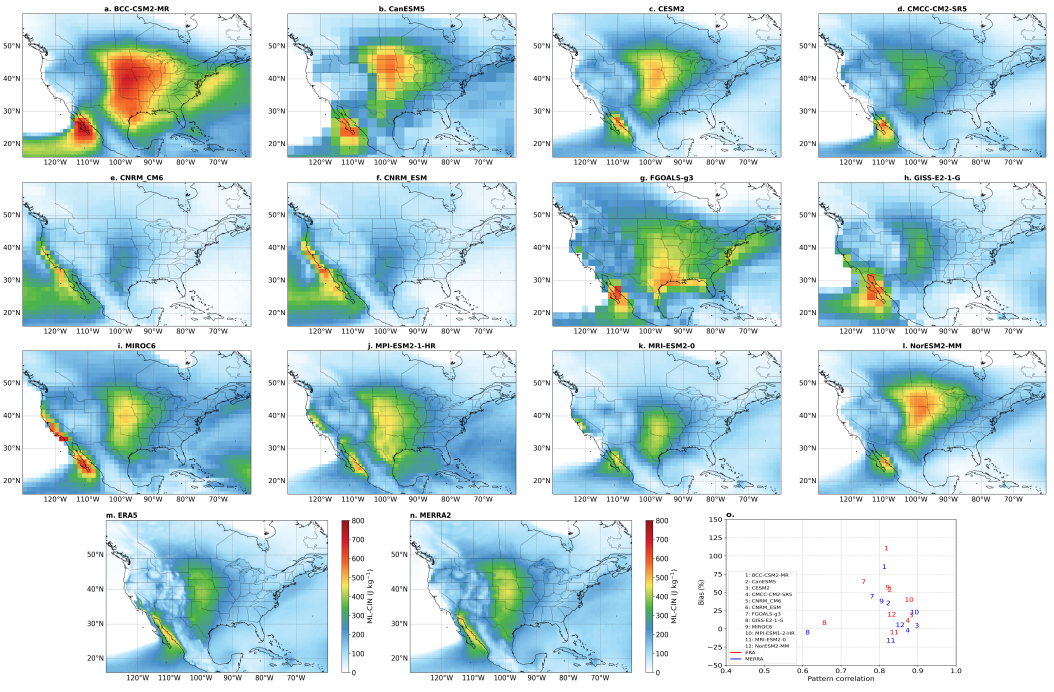

FIG. 7. As for Fig. 6 but for ML-CIN. In (o), CNRM-CM6 and CNRM-ESM do not appear as their pattern correlation is about 0.3 and 0.2, respectively.

models with much higher S06 values near the northeast US. MPI-ESM1.2-HR and MRI-ESM2 show relatively good skill in capturing the S06 spatial structure seen in the reanalyses (close to 0 bias and >0.92 pattern correlation). Although FGOALS-g3 shows a better skill in capturing the S06 pattern in the northeast region, it fails to reproduce the high shear values in the northwest US (0.87 pattern correlation). CMCC-CM2-SR5 and the two CNRM models underestimate the S06 magnitude by ~10%. As in spring, the intermodel spread in terms of relative range of S06 magnitude is the smallest among the four convective variables and is 0.18.

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

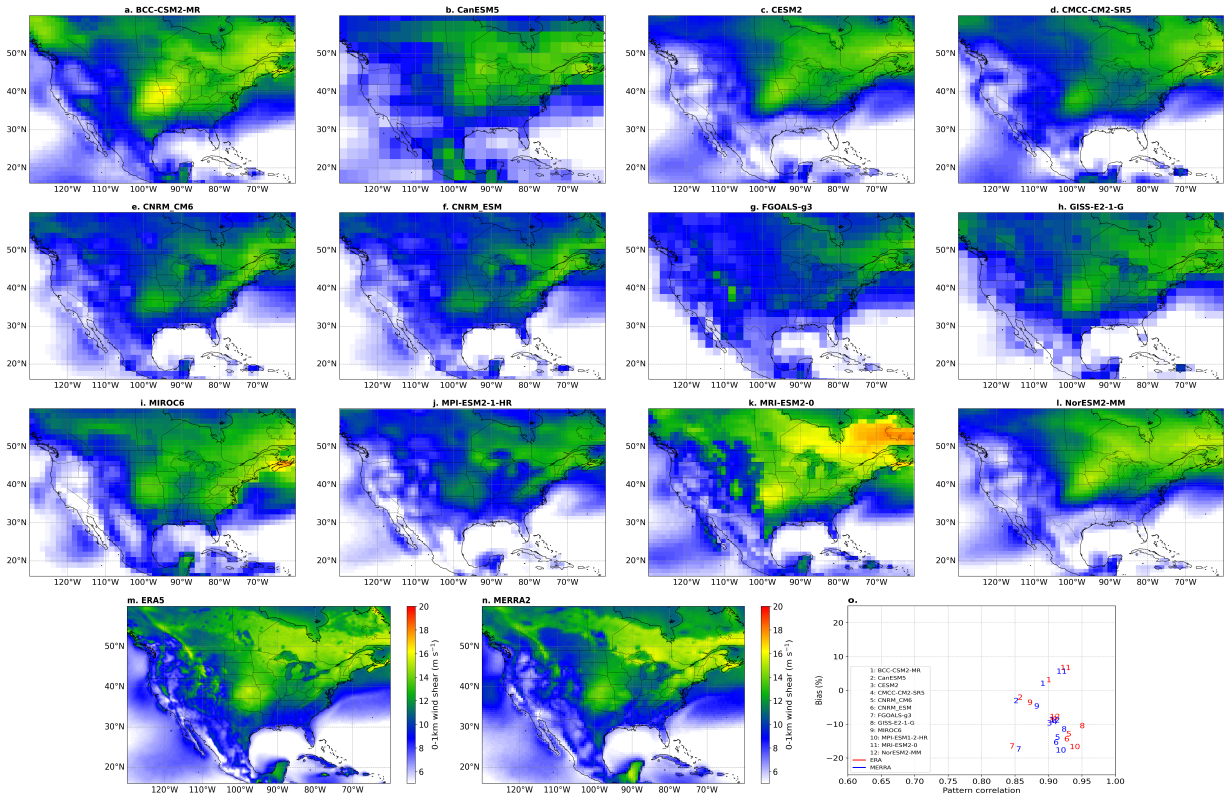

FIG. 8. As for Fig. 6 but for S01.

3) FALL (SEPTEMBER–NOVEMBER) AND WINTER (DECEMEBER–FEBRUARY)

Fall P95 CAPE values are predominantly higher over oceanic areas, with peak reanalysis values around 2000 J kg^{-1} (Figure SF6). Models show higher values (close to 4000 J kg^{-1}) over GoM and the Caribbean Sea. In contrast, the pattern of P95 ML-CIN is well captured (Fig. SF7). Though models slightly underestimate the P95 S01 pattern (Fig. SF8), most models fail to capture the P95 S06 field (Fig. SF9) accurately. The relative range for CAPE, CIN, S01, and S06 in fall is 2.45, 0.90, 0.26, 0.15, respectively. P95 ML-CAPE in winter (Fig. SF10) is very weak over the region with values well below 1000 J kg^{-1} over the Caribbean Sea. A few models heavily overestimate the oceanic P95 ML-CAPE values. Wintertime P95 ML-CIN is lower in most models (Fig. SF11). Except BCC-CSM2-MR and MPI-ESM1.2-HR, all other models considerably underestimate the P95 ML-CIN magnitude. MRI-ESM2 performs best among the CMIP ensemble in capturing P95 S01 (Fig. SF12), whereas CESM2, FGOALS-g3, and NorESM2-MM have lower P95 S01 magnitudes. While models show a reasonable skill in capturing P95 S06, most models have a

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

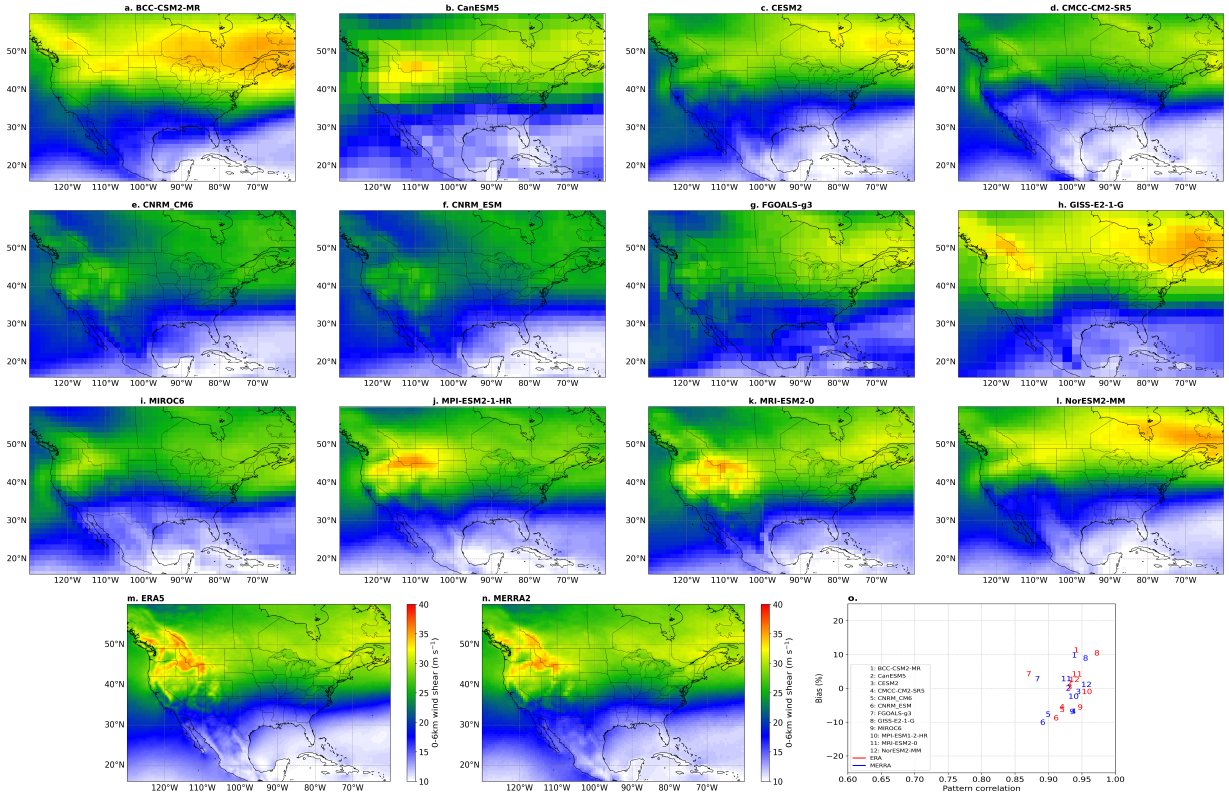

FIG. 9. As for Fig. 6 but for S06.

smoother pattern with a slightly lower magnitudes. The relative range for wintertime CAPE, CIN, S01, and S06 is 2.51, 1.16, 0.24, 0.19, respectively.

The analysis of convective parameters suggests that while most models have a tendency to overestimate ML-CAPE, wind shear is lower in the majority of the models. More importantly, several models have similar spatial patterns, suggesting a basis for categorizing them into specific groupings. Based on ML-CAPE patterns, CMCC-CM2-SR5, CanESM5, GISS-E2.1-G, and MIROC6 can be categorized as a 'high-CAPE' family of models. The rest of the models may fall into a 'moderate-CAPE' family. The intermodel spread quantified in terms of relative range of parameter magnitude suggests that CAPE has largest model spread and S06 has smallest model spread in all four seasons.

c. Mean spatial patterns

The parameter distributions (Fig. 1) and the spatial distribution of extremes (P95) of convective parameters revealed varying degrees of biases. To assess models' skill in reproducing seasonal

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

mean (climatology) of each parameter, we compute pattern correlation and standard deviation of seasonal mean relative to the reanalyses. Since each models and the two reanalyses operate on different native grids, to avoid spatially induced parameter biases, all datasets are regridded to a uniform 1°×1°mesh to facilitate a grid-to-grid comparison. The pattern correlation coefficient and standard deviation are plotted as Taylor diagrams for all four parameters. For a consistent comparison, we normalize the standard deviation of each parameter with the reference standard deviation (computed from the reanalysis). Normalization facilitates easier comparison, and a normalized standard deviation (NSD) close to 1 indicates that the model captures the spatial variability well.

1) Spring

The models in general show larger spatial variability for ML-CAPE when compared to the reanalysis climatology (Fig. 10a) as evidenced by large NSD. The skill scores for the models are slightly different between ERA5 and MERRA2, primarily in terms of NSD. The spatial pattern is well captured by most models with a pattern correlation close to 0.9 consistent with earlier findings (Chavas and Li 2022). CanESM5 and FGOALS-g3 show relatively weaker correlation (~0.8). GISS-E2.1-G, CMCC-CM2-SR5, CNRM-CM6, CNRM-ESM, FGOALS-g3, and MIROC6, exhibit a very large NSD. The intermodel spread in terms of the relative range of climatological CAPE magnitude is 1.9. Note that relative range for seasonal mean is larger than that of P95.

The intermodel spread in terms of ML-CIN NSD (Fig. 10b) is much lower as compared to that of CAPE. However, all models except BCC-CSM2-MR under-represent the spatial variability (NSD) relative to the reanalysis data. NorESM2-MM reproduces only half of the spatial variability relative to reanalysis. As observed for P95 CIN, this models have a considerable negative bias in reproducing the CIN magnitude as well. MPI-ESM-1.2-HR captures the spatial pattern and the variability well with a correlation close to 0.95. MIROC6 also shows a good skill, with a pattern correlation of 0.9. All other models underestimate the spatial variability (except BCC-CSM2-MR), though the pattern correlation is above 0.85. The relative range in terms of CIN magnitude is 1.19, suggesting large ensemble spread, and the spread is much larger as opposed to that of P95 values (0.87).

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

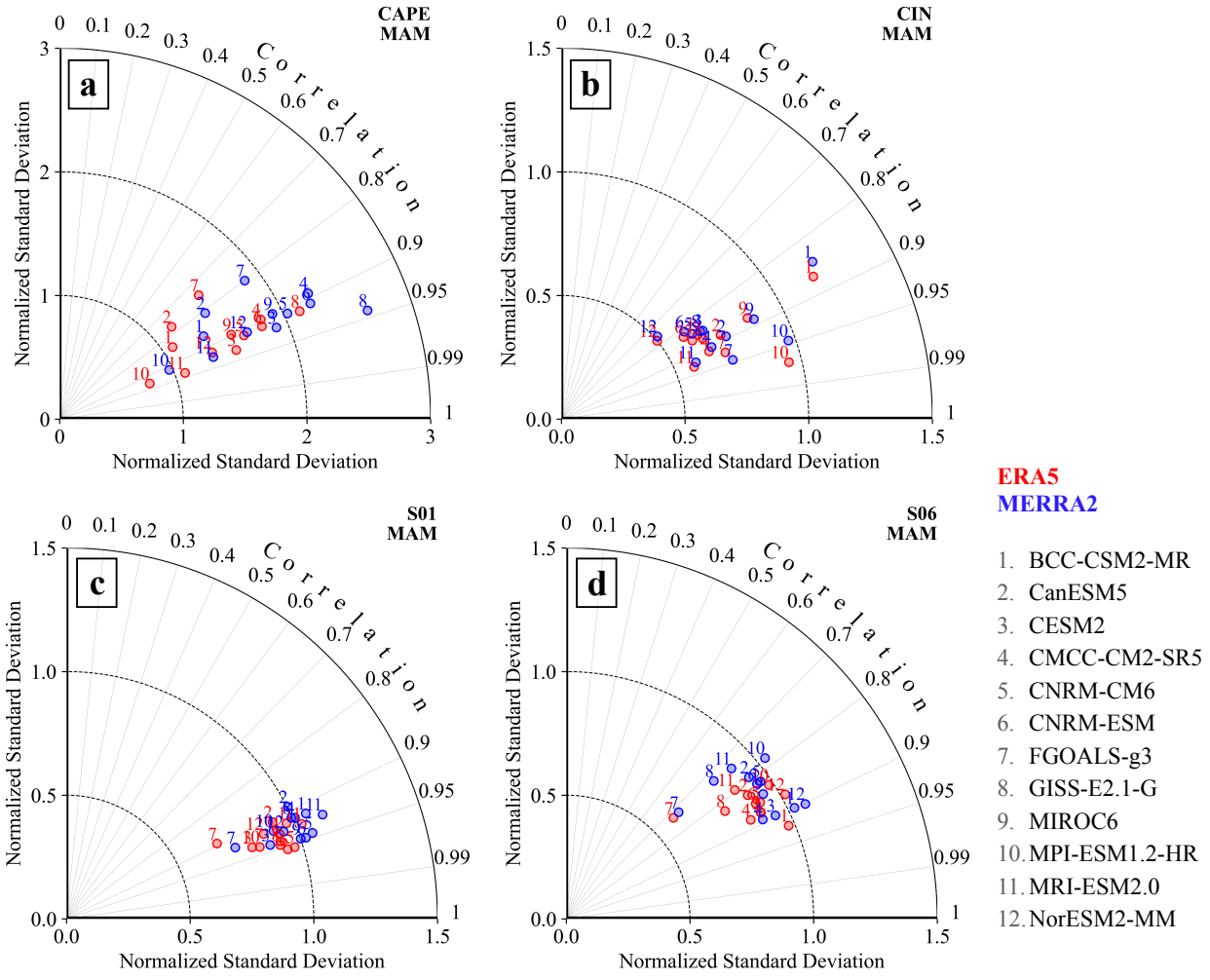

FIG. 10. Taylor diagram showing pattern correlation and normalized standard deviation for seasonal climatologies of (a) ML-CAPE and (b) ML-CIN (c) S01, and (d) S06 for spring. The pattern correlation and standard deviation are computed for each model with respect to ERA5 (red) and MERRA2 (blue). Note that BCC-CSM2-MR does not appear in panel d as it lies out of the quadrant due to high NSD (2.07).

The S01 mean pattern in spring (Fig. 10c) is reasonably well captured by most models (>0.9 pattern correlation) and there is a better agreement among the models as compared to the thermodynamic parameters. However, NSD is underestimated by FGOALS-g3, CESM2, NorESM2-MM, MPI-ESM1.2-HR. Models' skill in reproducing the mean S06 (Fig. 10d) pattern is slightly weaker than that of S01. A consistent outlier is FGOALS-g3, which shows a pattern correlation of ~0.7 with under-represented spatial variability (by almost 40%). BCC-CSM2-MR, CESM2, NorESM2-

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

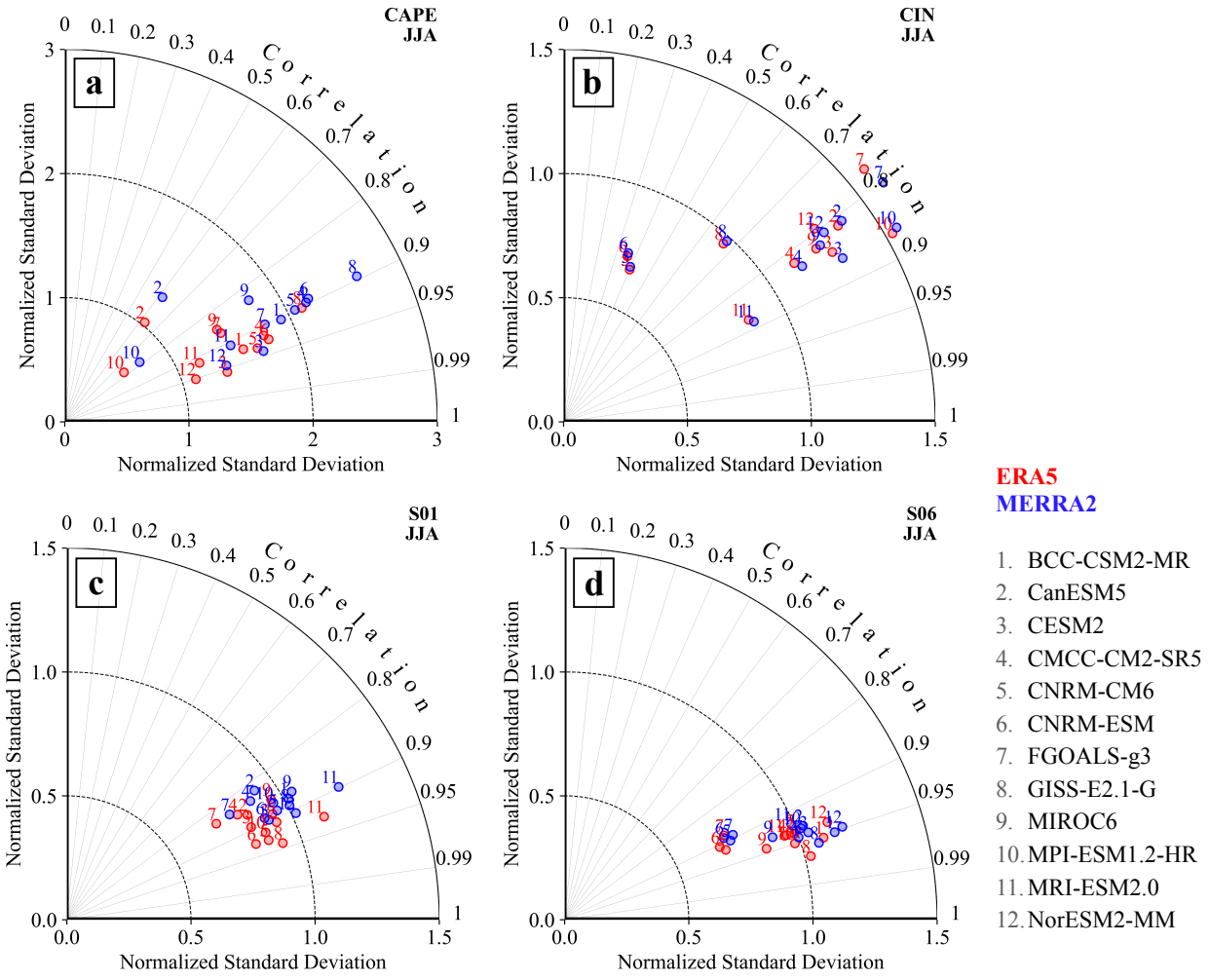

FIG. 11. As for Fig. 10 but for summer.

MM, and CMCC-CM2-SR5 show relatively better skill. The relative range for climatological S01 and S06 magnitudes is (0.30 and 0.17) very similar to that of the P95 values (0.29, and 0.17).

2) SUMMER

The intermodel spread in summertime ML-CAPE in terms of pattern correlation and NSD (Fig. 11a) is larger when compared to spring. The pattern correlation for CanESM5 is notably weaker (~0.6) compared to other models, possibly due to the presence of the extended CAPE maxima in CanESM5 (Fig. 6). Curiously, MPI-ESM1.2-HR shows relatively lower pattern correlation (approximately 0.8) with smaller NSD. GISS-E2.1-G, CMCC-CM2-SR5, CNRM-CM6,

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

and CNRM-ESM show better correlation, albeit with larger NSD. The relative range of CAPE magnitude is 2.51, and this also underscores a high ensemble spread.

The summertime CIN climatology (Fig. 11b) also shows a larger spread than that in spring. A similar spread is seen in relative range based on CIN magnitude (2.0), which is also larger as compared to the relative range in P95 CIN magnitude (1.3). Both the CNRM models show much weaker pattern correlation (0.4) with respect to the reanalysis climatology. Another outlier is BCC-CSM2-MR, which has significantly large spatial variability with an NSD of 2.07 (lies out of the quadrant), possibly due to its the large positive bias. GISS-E-2.1-G captures the spatial variability quite well, however the pattern correlation is relatively lower (~0.65). All other models show a pattern correlation higher than 0.8. Among them, CESM2, MRI-ESM2.0, and CMCC-CM2-SR5 show NSD close to 1.

Most models capture the S01 mean climatology in summer (Fig. 11c) reasonably well. Note that the NSD values for the models are slightly different with respect to ERA5 and MERRA2, with values corresponding to MERRA2 are larger. MRI-ESM2.0 overestimates the spatial variability (NSD ~1.2), though it captures the pattern well (0.9 correlation). NSD with respect to ERA5 suggests that many other models tend to underplay the spatial variability, specifically FGOALS-g3 (NSD ~0.75). The relative range of S01 magnitude is 0.3. In terms of S06 climatology (Fig. 11d), most models reproduce the spatial variability and pattern well. The S06 relative range is 0.24, which is larger compared to spring (0.17). The CNRM models and FGOALS-g3 underestimate the spatial variability, with an NSD of ~0.6. FGOALS-g3 consistently underestimates the spatial variability of S01 and S06, in both seasons.

3) Fall and Winter

The ML-CAPE climatology in fall (Fig. SF14a) also exhibits a large intermodel spread. Models, except CanESM5 show a pattern correlation higher than 0.8. The ML-CIN climatology from models (Fig. SF14b) also displays large spread. Several models struggle to reproduce the spatial pattern, specifically the CNRM models. The S01 climatology in fall (Fig. SF14c) is fairly well captured by most models, except FGOALS-g3, CESM2, NorESM2-MM, and CanESM5. The results are mostly similar for S06 climatology (Fig. SF14d). In winter, GISS-E1.2-G, FGOALS-g3, CMCC-CM2-SR5, CNRM-CM6, and CNRM-ESM overestimate the oceanic CAPE though they

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

capture the pattern fairly well (Fig. SF15a). MPI-ESM1.2-HR considerably underestimates the spatial variability. While models capture the mean ML-CIN (Fig. SF15b) spatial pattern quite well, except BCC-CSM2-MR and the MPI model, all other models underestimate the spatial variability. The mean S01 patterns in models show better correlation score with MERRA2 (Fig. SF15c) as compared to ERA5. FGOALS-g3, CESM2, and NorESM2-MM underestimate the spatial variability, though the pattern correlation is fairly high (>0.9). Similar to other seasons, most models tend to underestimate the spatial variability of S06 climatology (Fig. SF15d), particularly FGOALS-g3 (NSD ~0.6).

d. Regional biases in seasonal means

The analyses in previous sections reveal that the biases in convective parameters are spatially non-homogeneous and vary across the models. Here, we examine and quantify the seasonal mean biases in CAPE, CIN, S01, and S06 during spring and summer across three subdomains over the continental US: (i) western CONUS that lies west of the Rocky Mountains, (ii) central CONUS, which comprise of much of the Great Plains, and (iii) eastern CONUS to the east of the Great Plains (shown in SF16).

Most models overestimate the mean CAPE over the western region (Fig. 12a,b) where climatological values are low. CAPE bias is lowest in the central CONUS for most models, however CanESM5, CMC-CM2-SR5, GISS-E2.1-G, and MIROC6 have higher CAPE values relative to ERA5. FGOALS-g3 is the only model that shows slight underestimation of springtime CAPE. BCC-CSM2-MR, CESM2, MPI-ESM1.2-HR, and MRI-ESM2.0 are better in terms of capturing the mean CAPE fields in both seasons. In the eastern subdomain, models tend to overestimate CAPE, with FGOALS-g3 and MPI-ESM1.2-HR performing better. Bias in mean CIN (Fig. 12c,d) is also largest over the western CONUS, especially in spring. Over the central CONUS, models tend to underrepresent CIN, especially in spring. The two CNRM models and MRI-ESM2.0 exhibit almost 50% reduction in the mean CIN in both seasons, whereas the BCC-CSM2-MR exhibits considerable overestimation (~90%) during summer. BCC-CSM2-MR, CanESM5, and MIROC6 overestimate the springtime CIN values in eastern CONUS, whereas in summer, BCC-CSM2-MR and FGOALS-g3 exhibit very high CIN values. The CNRM models, GISS-E2.1-G,

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

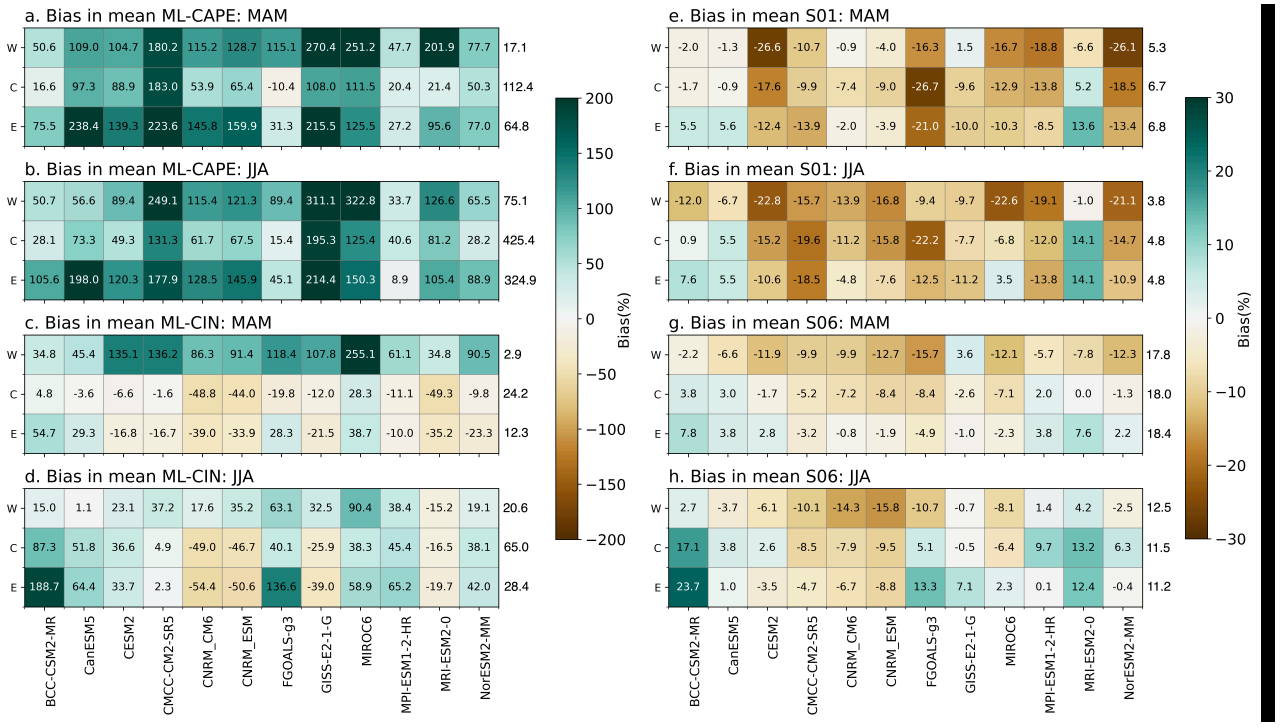

FIG. 12. Heatmaps showing mean fractional biases in (a,b) CAPE, (c,d) CIN, (e,f) S01, and (g,h) S06 during (a,c,e,g) spring and (b,d,f,h) summer for the three subdomains across the CONUS. W, C, and E on the y-axis are for Western, Central, and Eastern subdomains. Fractional bias for models is computed using 6h data on 1×1 lat-lon mesh, relative to ERA5-derived fields. The numbers on the right side of heatmaps indicate mean values of each parameter from ERA5 corresponding to each subdomain.

and MRI-ESM2.0 consistently underestimate mean CIN over central and eastern CONUS in both the seasons.

Most models underestimate S01 throughout the CONUS, however the magnitude of underestimation lies within 20% in general (Fig. 12e,f). In spring, CESM2, FGOALS-g3, and NorESM2-MM show large difference in mean S01 relative to ERA5. BCC-CSM2-MR, CanESM5, the two CNRM models, GISS-E2.1-G, and MRI-ESM2.0 have relatively lower bias over all three sub-domains in spring. Whereas, over the central CONUS, CESM2, FGOALS-g3, and NorESM2-MM exhibit large deviation from ERA5. The underestimation becomes more evident in summer in many models. CESM2, MIROC6, and NorESM2-MM show more than 20% reduction in mean S01 values. All the models invariably underestimate mean S01 over the western CONUS in the summer season. The pattern is almost similar in the eastern sub-domain as well. Most models underestimate the

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

S06 magnitude also across CONUS (Fig. 12g,h). Over the central and eastern CONUS, models show a good skill in reproducing S06 magnitude with mean bias within 10% in spring. Over the central CONUS, the BCC-CSM2-MR and MRI-ESM2.0 overestimate S06. While most models show relatively better skill over the eastern region, BCC-CSM2-MR exhibits a large positive bias (+23%).

Bias in convective parameters vary considerably across the continental US. Most models struggle to accurately simulate the parameter magnitude over the western CONUS, which is likely to be associated with resolving topography near the Rockies. The bias values are relatively lower in the central CONUS, where the severe thunderstorm frequency is highest but are larger over the eastern US.

e. Bias in thresholded frequency

Mean biases in parameter magnitudes do not necessarily convey the potential for frequential biases in representing favorable severe thunderstorm environments. Here we discuss the bias in frequencies of ML-CAPE exceeding 1000 J kg^{-1}, ML-CIN exceeding 150 J kg^{-1}, S01 and S06 exceeding 10 m s^{-1} and 20 m s^{-1}, respectively.

Fig. 13a,b reveals that the bias in thresholded frequency is much larger than the bias in mean ML-CAPE. Most models have bias higher than 300% over the western and eastern CONUS in spring, which suggests that prevalence of event is not the sole driver of these regional signals. The frequency bias is relatively lower over the central region as compared to the other two sub-domains, though positively biased for most models. Interestingly, FGOALS-g3 shows a negative bias in ML-CAPE frequency. The frequency bias in ML-CAPE in summer is slightly smaller as compared to spring. In spring, models except MPI-ESM1.2-HR have a bias larger than 75% over the eastern CONUS, whereas in summer, the frequency bias is above 100% for most models with an exception of FGOALS-g3 and MPI-ESM1.2-HR. The frequency bias for ML-CIN exceeding 150 J kg^{-1} is large in general (Fig. 13c,d). Over the western CONUS, ML-CIN frequency is predominantly positive and higher than 50% in spring. However, in the central CONUS, most models exhibit a negative bias, and interestingly, they show varying patterns for CIN frequency in the eastern region during spring. The CNRM models consistently show negative bias (more than 60%) over the central CONUS in spring and summer. In the summer season BCC-CSM2-MR,

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

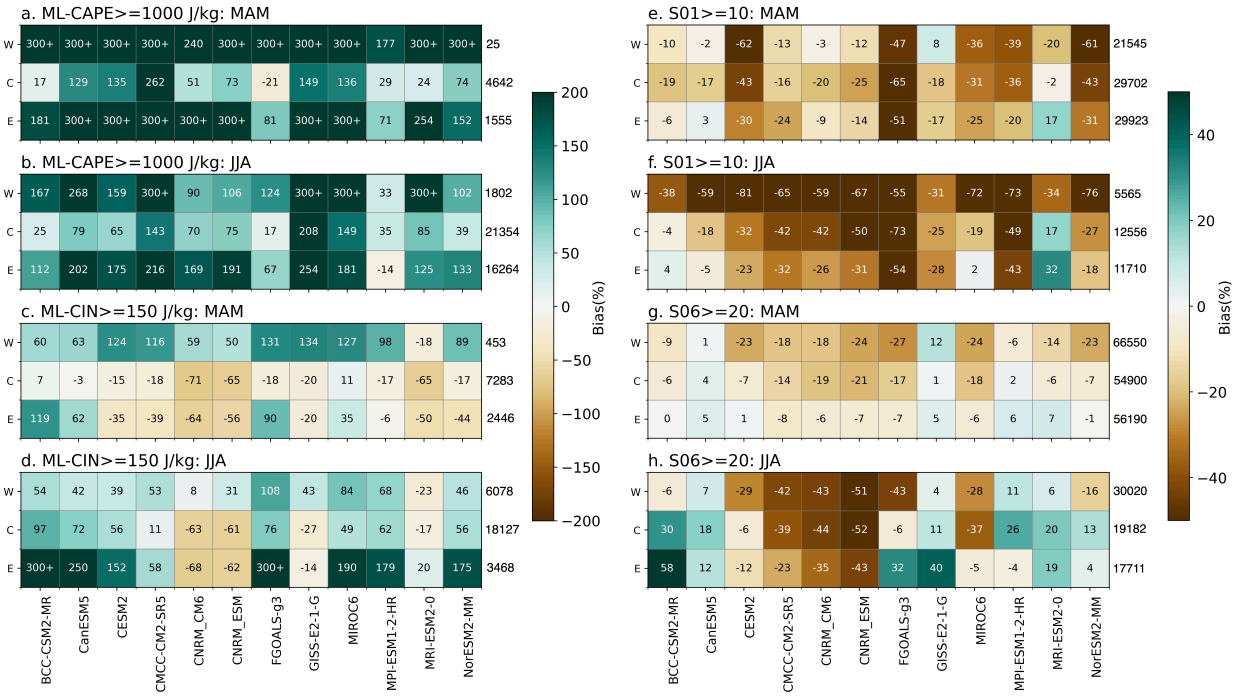

FIG. 13. Heatmaps showing frequency biases in (a,b) CAPE, (c,d) CIN, (e,f) S01, and (g,h) S06 in (a,c,e,g) spring and (b,d,f,h) summer for the three subdomains across the CONUS. W, C, and E on the y-axis are for Western, Central, and Eastern subdomains. Fractional bias is computed for 6h data at 1×1 lat-lon grid with respect to ERA5-derived fields. The numbers on the right of heatmaps show the total number of events satisfying the condition in each subdomains from ERA5 data.

CanESM5, CESM2, FGOALS-g3, MIROC6, MPI-ESM1.2-HR, and NorESM2-MM overestimate the CIN frequency by almost 50% or higher in the central region. Over the eastern region in summer, in contrast to spring, most models show much larger bias in CIN frequency.

The bias in thresholded frequencies for S01 and S06 (Fig. 13e-h) are predominantly negative across the US. More importantly, as in the case of convective parameters, the frequency bias is considerably larger as compared to the bias in the mean. CESM2, FGOALS-g3, and NorESM2-MM significantly underestimate the frequency S01 exceeding 10 m s^{-1} in all three subdomains in spring. The severity of negative bias increases in most models in the summer season, except MRI-ESM2.0. FGOALS-g3 and MPI-ESM1.2-HR exhibit large underestimation in all three subdomains in summer. The bias in frequency of S06 exceeding 20 m s^{-1} (Fig. 13g,h) is modest in spring when compared to the bias in S01 frequency. Negative bias is largest over the western CONUS for several models. In general, underestimation in S06 frequency is within 20% over the central

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

CONUS region and within 5% over the eastern region. However in summer, the above pattern changes and several models show substantial negative bias.

We analyzed the thresholded frequency bias in all four parameter for higher thresholds as well and the bias is generally higher (Fig. SF17).

To examine how the bias in component variables (CAPE, CIN, and S06) impact the likelihood of severe storm formation, we calculated supercell composite parameter (SCP) following the definition in Lepore et al. (2021a). We computed the bias in frequency of events where SCP exceeds 1.0 in each sub-domain for spring and summer (Fig. SF18). CanESM5, GISS-E2.1-G, MIROC6, and MRI-ESM2.0 exhibit large positive bias in SCP frequency in both seasons across the US (140 to 300+ % bias). As we noted in other convective variables, the bias is generally lower in the central CONUS. BCC-CSM2-MR, CESM2, FGOALS-g3, MPI-ESM1-2-HR, and NorESM2-MM show bias within +/- 50% over central CONUS.

f. Bias in vertical structure

The results discussed so far indicate varying degrees of biases in convective parameters from CMIP6 models. It is very likely that these biases are reflective of the biases in the basic meteorological fields. To validate this hypothesis, we examine the vertical profiles of model-simulated temperature, specific humidity, and wind fields with respect to the observed data for the three subdomains defined in the previous sections.

The 50^{th} percentile of temperature bias distributions in spring for W. CONUS (Fig. 14a) indicate that all models have a noticeable warm bias throughout the lower and middle troposphere. GISS-E2.1-G, both CNRM models, and FGOALS-g3 show larger warm bias. These models exhibit largest warm bias in summer as well (Fig. 14b). The CNRM models and GISS-E2.1-G have biases up to 10 K near 9 km in vertical. CMCC-CM2-SR5 and NorESM2-MM exhibit a summertime cold bias (-2 K) near the surface. MPI-ESM1.2-HR, CMCC-CM2-SR5, CanESM5, NorESM2-MM show relatively smaller temperature bias (2 K within ~10km) in summer. The bias in springtime specific humidity (Fig. 14c) shows mixed signals within the lower 8km. CanESM5 and GISS-E2.1-G has the largest positive bias in the boundary layer (0.6 g kg^{-1}), whereas, MRI-ESM2.0 has a dry bias of 1 g kg^{-1} near the surface. Many other models also show a dry bias within the lower troposphere. This pattern reverses above 3km, and most models exhibit a moist bias up to 8km

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

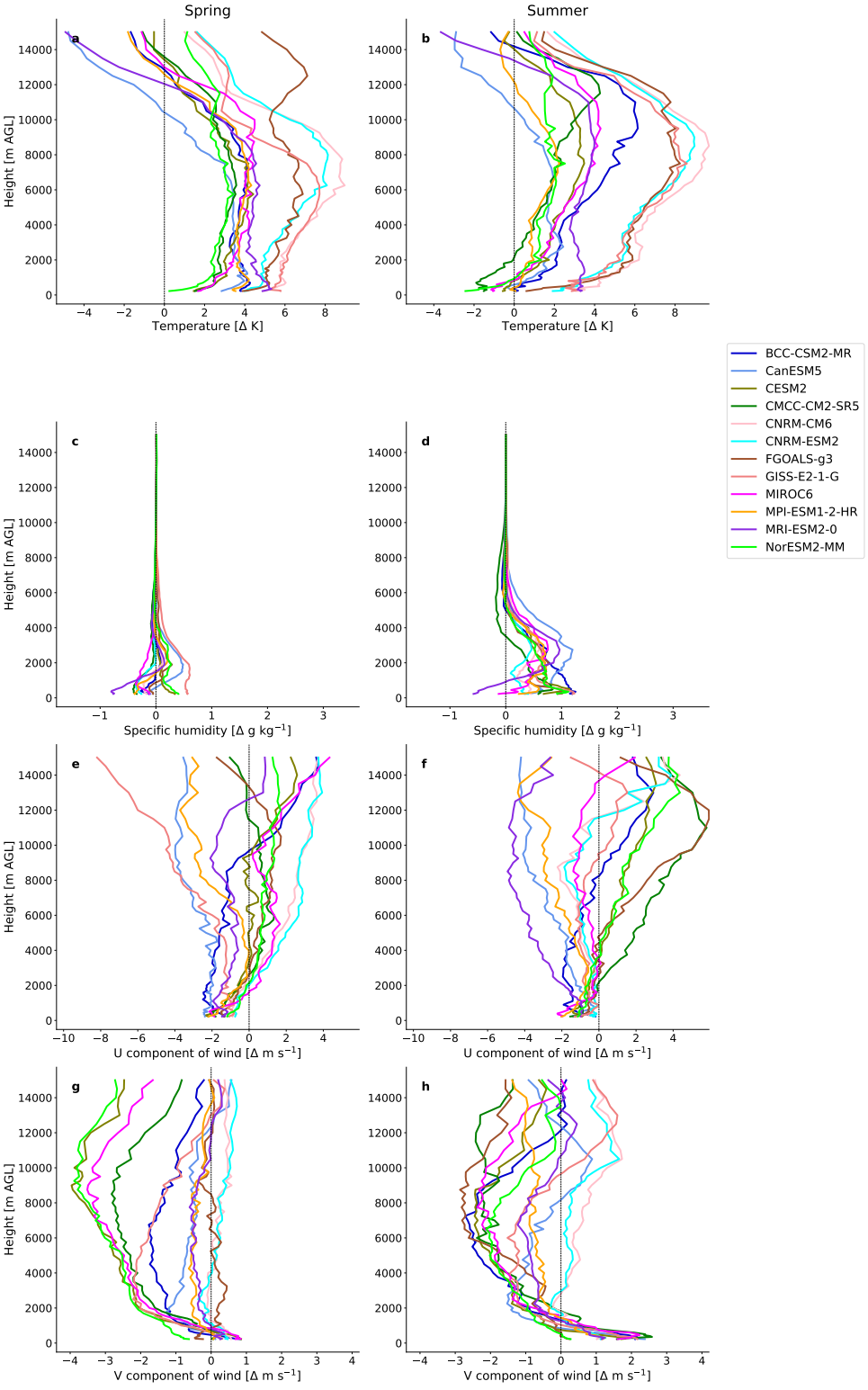

FIG. 14. 50th percentile of bias distributions in (a, b) temperature, (c, d) specific humidity, (e, f) u-wind, and (g, h) v-wind in models with respect to IGRA2 observations on interpolated height levels for W. CONUS subdomain.

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

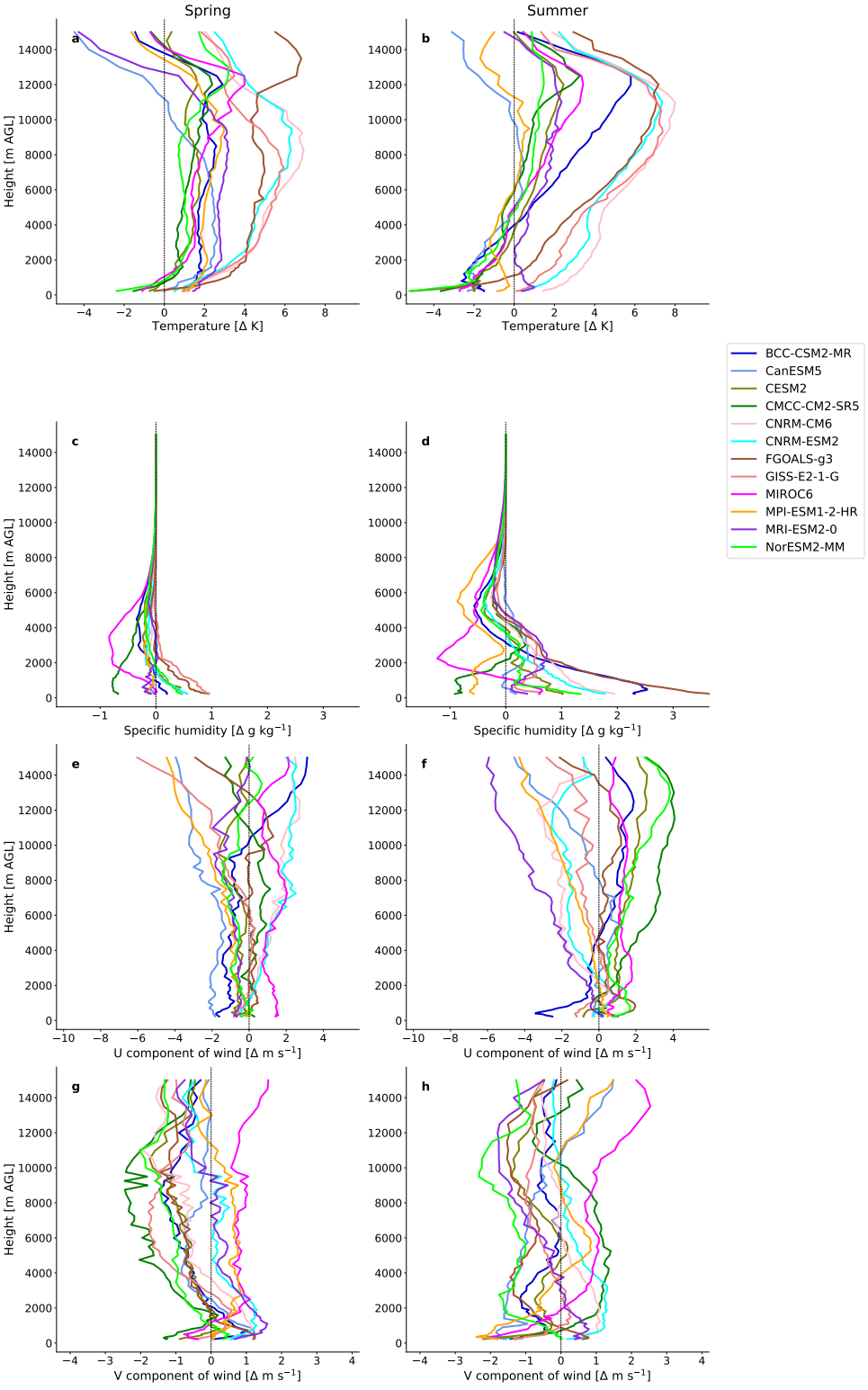

FIG. 15. As for 14 but for C. CONUS subdomain.

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

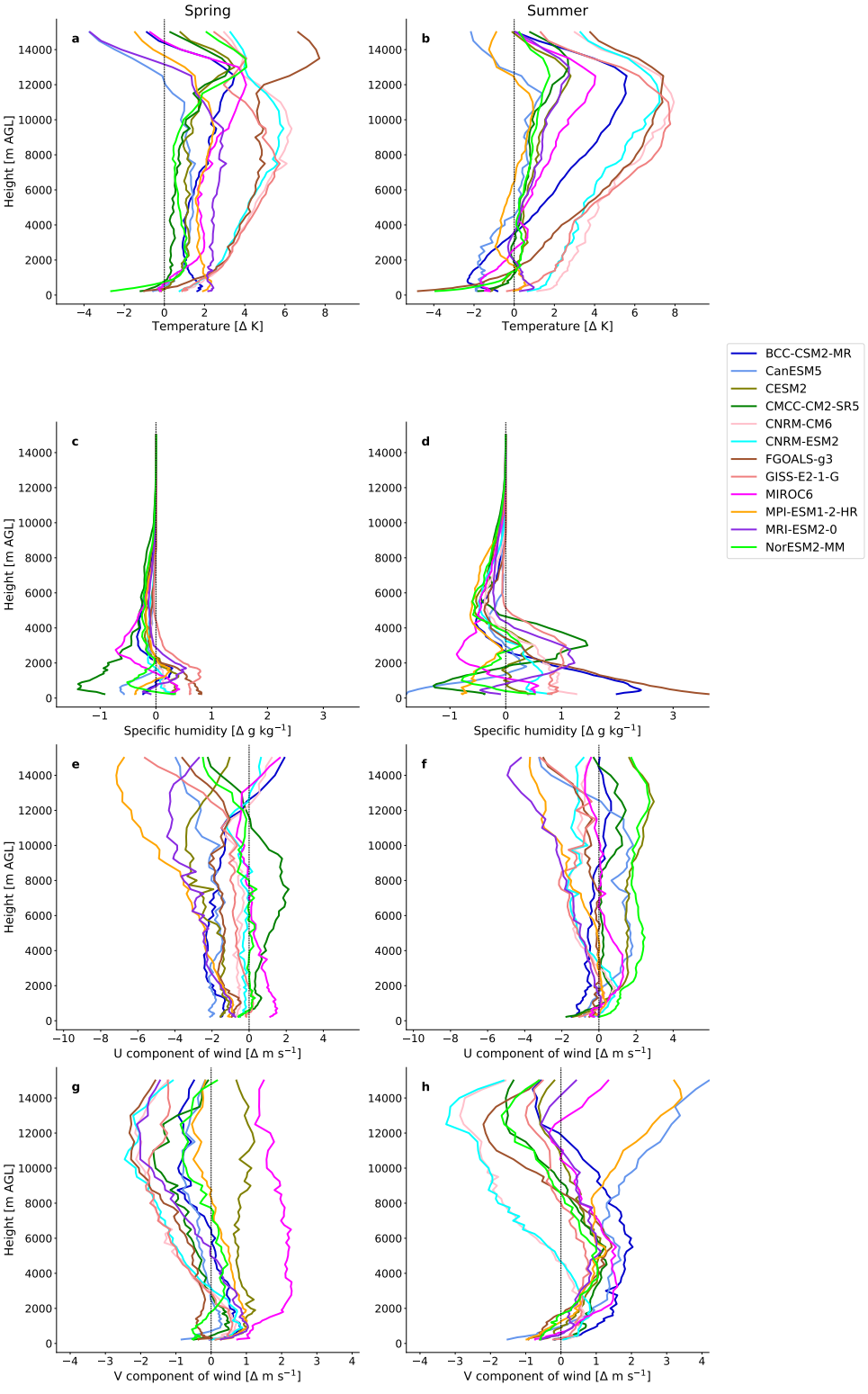

FIG. 16. As for 14 but for E. CONUS subdomain.

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

in vertical. The moist bias is relatively larger in summer (Fig. 14d), with almost all models have 0.5 g kg^{-1} moist bias within 0.5 – 3km in vertical. MRI-ESM2.0 shows a near-surface dry bias in summer as well. CanESM5 shows the largest moist bias near 3km. CMCC-CM2-SR5 exhibit positive bias within 3.5km and negative bias above through 10km.

All models underestimate the surface u-wind in spring and summer (Fig. 14e-h). Change in bias profile in vertical is indicative of bias in S01 and S06. BCC-CSM2-MR, CanESM5, and GISS-E2.1-G show relatively stable bias profile, albeit negatively biased. These models have been observed to have a better representation of wind shear. In spring, the CNRM models and MIROC6 exhibit largest changes in u-wind bias in vertical (0 – 6km). The above patter changes in summer, though most models have a negative bias. CMCC-CM2-SR5 and MRI-ESM2.0 show largest deviation in u-wind bias profile within 6km. The bias profiles for v-wind (Fig. 14g-h) indicate positive bias near-surface and negative bias above $\sim$ 1km for most models. While CNRM models, FGOALS-g3, MRI-ESM2.0 etc. show minimal changes in bias in vertical, NorESM2-MM, CESM2, MIROC6, and CMCC-CM2-SR5 exhibit the largest changes (up to 3 m s^{-1} at 6km height) in spring. The change in surface to 1km v-wind bias is more pronounced in summer (up to 3 m s^{-1} at 1km height). This will contribute to a bias in both S01 and S06, as would be expected from biases noted in previous sections.

Similar analysis for C. CONUS (Fig. 15) and E. CONUS (Fig. 16) confirms that the temperature bias in larger in the W. CONUS region. The CNRM models, FGOALS-g3, and GISS-E2.1-G have the largest temperature bias in all three sub-domains. Most models exhibit similar temperature bias profiles in all three regions, albeit with slightly different magnitudes. Whereas, the bias profiles for specific humidity in the C.CONUS (Fig. 15c-d) suggest significant dry bias in MIROC6, specifically in the lower- to mid-tropospheric region (up to 1 g kg^{-1} at 2–4km). MPI-ESM1.2-HR also shows a dry bias in summer. FGOALS-g3 and BCC-CSM2-MR exhibit an excessively moist bias in the lower troposphere, particularly in summer. Models tend to show larger moisture bias in summer as compared to spring in the lower troposphere. The above results are more or less similar for the E.CONUS as well (Fig. 16c-d). The bias in u-wind over C.CONUS indicates that near-surface wind bias is somewhat lower relative to that in W.CONUS. Additionally, changes in bias profile is lower in spring as compared to summer, corroborating earlier findings that bias in S01 and S06 are larger in summer. However, u-wind profiles have similar bias in spring and summer in the

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

E.CONUS sub-domain. Most models have a negative u-wind bias in spring. The bias in v-wind over the C.CONUS (Fig. 15e-f) shows more changes in vertical, suggesting the potential for bias in both S01 and S06. MIROC6, MPI-ESM1.2-HR (in summer), CMCC-CM2-SR5 etc. exhibit the largest changes in bias in the vertical. In the E.CONUS sub-domain, v-wind bias (Fig. 16e-f) is largest in the CNRM models, FGOALS-g3 and MPI-ESM1.2-HR (in summer).

The positive bias in temperature and moisture fields observed in the majority of the models tends to corroborate the excessive bias in thermodynamic parameters identified in the preceding sections. The two interesting cases are of FGOALS-g3 and BCC-CSM2-MR (in summer, except in W.CONUS). Among the above-mentioned models, FGOALS-g3 exhibits a large warm and moist bias; however, it shows only a moderate magnitude of CAPE. More intriguingly, this model has a minor underestimation in ML-CAPE over the central CONUS region. A possible reason for such a behavior can be that the biases may be predominant on non-convective days. Similarly, BCC-CSM2-MR also shows a substantially high humidity bias in summer. Curiously, mean CIN during summer is also higher, specifically over central and eastern CONUS. On the other hand, the two CNRM models display consistently lower CIN values over central and eastern CONUS. This can be attributed to the fact that these two models have a low-level warm moist bias, which may result in smaller CIN values. However, analyzing bias in relative humidity profiles (Fig. SF19), we find that most models have a negative bias in the lower- to mid-troposphere. Models' tendency to underestimate extremes in CIN (P95) could be associated with the negative bias in RH. The high CAPE values in models such as GISS-E2.1-G and CanESM5 may also be attributed to the considerable warm bias and positive humidity bias present in these two models. Interestingly, CESM2 and NorESM2-MM exhibit very similar temperature bias profiles, although there are slight differences in their humidity bias profiles. MPI-ESM1.2-HR and MRI-ESM2.0 show only a moderate bias in temperature and humidity, though the springtime humidity bias is quite different in these models. Results were cross-validated using the ERA-5 and MERRA2 datasets, and showed similar persistent bias structures suggesting that evaluating against only the reanalyses may hinder identification of such biases (not shown).

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

4. Discussion and conclusions

The results presented here reveal varying degrees of biases in all thermodynamic and kinematic parameters in CMIP6 models relative to the reanalysis-derived parameters. Generally, CAPE is positively biased in the models, whereas the vertical wind shear is negatively biased. Previous studies that have examined the current-climate severe thunderstorm environment distributions in CMIP models (e.g., Diffenbaugh et al. 2013; Seeley and Romps 2015; Chavas and Li 2022) have also reported significant positive bias in CAPE. Chavas and Li (2022) reported that MPI-ESM1.2-HR, CNRM-CM6, and CNRM-ESM as good models in terms of CAPE bias. Our study revealed that while upper quantiles of CAPE is well captured by MPI-ESM1.2-HR, the lower percentiles are underestimated (Fig. 17). Additionally, MPI-ESM1.2-HR underestimate thresholded frequency in S01, specifically in the central CONUS region (up to 49%), where severe thunderstorm activity is highest. CNRM-ESM and CNRM-CM6 suffer from significant underestimation of CIN in central and eastern CONUS (56–71%), which can also impact the overall severe thunderstorm frequency. Comparing the results for CMIP5 models' skill in simulating CAPE and CIN from Diffenbaugh et al. (2013), models have improved in their performance in their CMIP6 version. Particularly, pattern correlation of CIN is improved considerably. The comparison of the vertical profiles of model-simulated temperature and humidity with respect to the observations suggests that there exists considerable warm and moist bias in the models. Lin et al. (2017) reported warm bias in CMIP5 models over the central CONUS, originating primarily as a result of reduced precipitation. Several CMIP6 models also reported to have warm bias over the US and adjoining oceanic regions (e.g., Wu et al. 2019; Seland et al. 2020; Swart et al. 2019). Excessive bias in temperature coupled with a positive bias in humidity contributes to more conditional instability, and hence more CAPE. Higher low-level moisture availability also leads to lower lifting condensation level (LCL), which in part can lead to lower values of CIN. Most models (with the exception of BCC-CSM2-MR and MIROC6) tend to simulate relatively lower values of CIN as opposed to the reanalysis fields. The recent study by Chavas and Li (2022) analyzing CMIP6 models argued that the positive bias in the model-simulated CAPE arises primarily from the biases in mean-state near-surface moist static energy. Our results indicate a peculiar ML-CAPE pattern in CanESM5 (extending excessively eastward) and CMCC-CM2-SR5 (a northwestward extension). Similar spatial patterns were noted in the 2-m specific humidity anomaly fields in these two models by Chavas and Li (2022), which

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

might explain the above-mentioned spatial bias in CAPE. Another potential reason for biases in CAPE and CIN is the errors arising from convective parameterization in the models as it can crucially impact CAPE dissipation (Trapp et al. 2007; Zhang 2009; Skinner and Diffenbaugh 2013; Allen et al. 2014b). In terms of CIN, a notable outlier is BCC-CSM2-MR, which tends to overestimate the CIN values. A possible reason would be that the LCL is set to nominal height of 650 hPa in this model (Wu et al. 2019) based on observed climatological values (Craven et al. 2004), which can be higher than the actual height of LCL. Somewhat unexpectedly, there does not appear to be a strong bias of thermodynamic or kinematic parameters on horizontal resolution in the climate models. For example, CanESM5 comes on a much coarser mesh with a grid spacing of approximately 2.8°. Though this model appeared to lack some finer details, it shows much better skill in simulating wind shear when compared to models such as CNRM-CM6, CNRM-ESM2, or MIROC6, which operate on higher horizontal resolution. On the other hand, despite having a relatively large number of vertical levels (18) within the lowest 500mb, CanESM5 exhibits a large thermodynamic bias. FGOALS-g3 has the least number of vertical grids (8) within the lowest 500mb among the ensemble considered here, which can potentially impact the capability of this model in resolving convective processes. Curiously, the CNRM models have the largest (27) number of vertical grids within 500mb, however exhibit significant low CIN bias.

We quantified the intermodel spread in terms of relative range of spatial mean magnitudes and found that the ensemble spread is largest for CAPE and smallest for S06 in all four seasons. Crucially, the bias for mean and extreme of convective parameters are considerably different, and therefore we analyzed how the bias vary across the distribution. The analysis reveals that the bias is not uniform for different percentiles (Fig. 17). More importantly, different models behave differently in terms of quantile biases. For the above reason, analyzing mean of parameters can also be misleading about the complete behaviour of the model. And the pattern varies seasonally as well. For instance, the springtime CAPE in GISS-E2.1-G is much overestimated at the lower percentiles as compared to the higher percentiles. On the other hand, MPI-ESM1.2-HR has large underestimation at the lower percentiles and the bias becomes near-zero towards P90 and above. The negative bias in CIN is more pronounced at lower percentiles in most models, and bias decreases towards P99. The quantile bias in S06 is notably different in spring and summer.

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

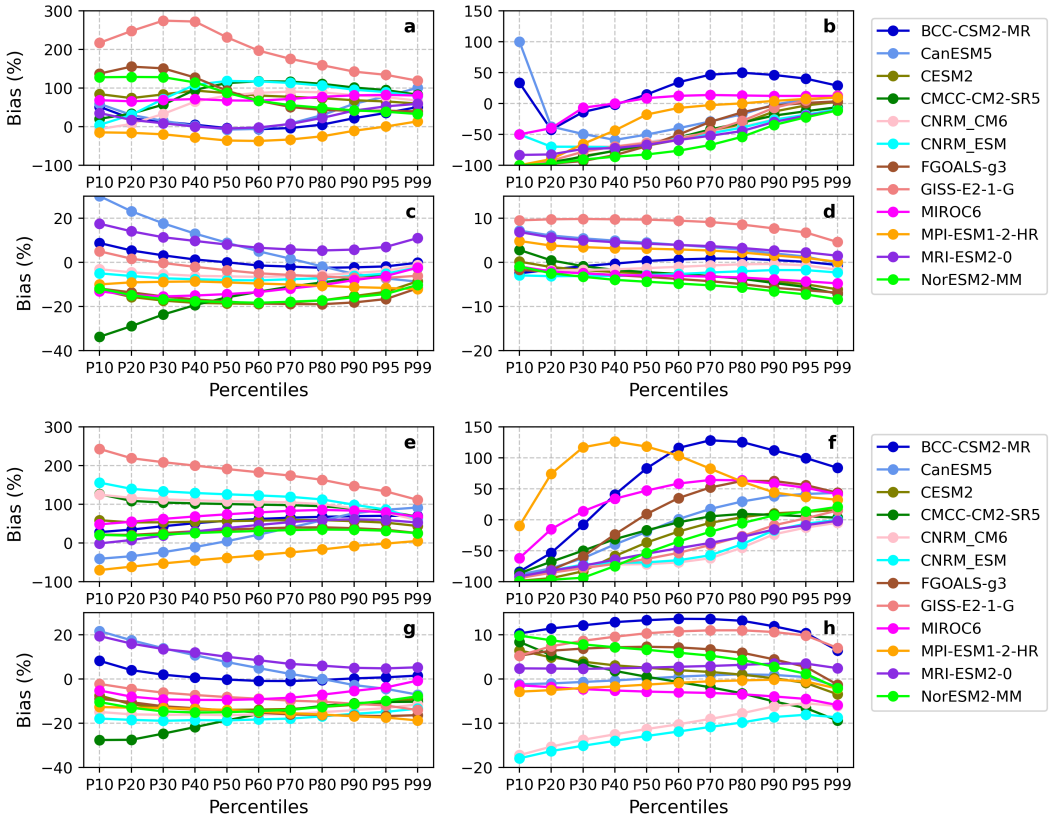

FIG. 17. Bias at different percentiles for (a,e) CAPE, (b,f) CIN, (c,g) S01, and S06 (d,h) S01 for (a-d) spring and (e-h) summer. Bias is computed for domain-averaged values of each parameter with relative to ERA5.

Biases related to oceanic convective 'hotspots' are another prevalent feature, with notably higher values of CAPE over the GoM and in the northern Atlantic (along the Gulf Stream). Most of the previous studies have overlooked this aspect. However, the study by Marsh et al. (2007) using CCSM3 indicated very similar positive CAPE bias over the oceans, particularly in spring. Even after several upgrades in coupled modeling systems, we note exactly similar convective hotspots along the Gulf Stream region in several models. This may be related to the simulation of sea surface temperatures (SST) and low-level humidity patterns or issues in the rendition of boundary layer fluxes or mixing in these models. For example, CMCC-CM2-SR5 and CanESM5 are shown to have a mean positive SST bias over these regions (Cherchi et al. 2019; Swart et al. 2019). BCC-CSM2-MR also reported to have a mean warm bias in near-surface air temperature over these oceanic regions, notably along the Gulf Stream (Wu et al. 2019). Multiple studies have shown the link between severe convective storm activity over North America and moisture source over the

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

GoM and North Atlantic area (Molina et al. 2018; Molina and Allen 2019, 2020). The excessive convective bias over these oceanic regions may advect inland and contribute to inland CAPE bias though it is likely there are also be continental contributions from the land-surface, particularly in the summer months (Molina and Allen 2019, 2020; Emanuel 2023; Tuckman et al. 2023). We have also found that the error statistics (bias, pattern correlation, and standard deviation in mean; from Taylor diagrams) improve if the oceanic regions are masked out.

Another notable finding is that there appears to be a familial behavior among the CMIP6 models analyzed here. In terms of ML-CAPE, CMCC-CM2-SR5, CanESM5, GISS-E2.1-G, and MIROC6 are the models falling into a 'high-CAPE' family of models that would seem to span model dynamics. The rest of the models group themselves into a 'moderate-CAPE' family. CMCC-CM2-SR5, CanESM5, CESM2, FGOALS-g3, and NorESM2-MM use modified versions of Zhang-McFarlane scheme to represent cumulus convection. However, these models differ in their CAPE and CIN biases, suggesting that models' thermodynamic bias is not solely dependent on the convective parameterization. Additionally, the community atmosphere model (CAM) is used as the atmospheric component in CESM2 (CAM6), CMCC-CM2-SR5 (CAM5), and NorESM2-MM (CAM6 with modifications). Interestingly, CESM2 and NorESM2-MM show similar kinematic parameters, however CESM2 exhibits slightly larger thermodynamic bias. On the other hand, CMCC-CM2-SR5, which uses earlier version of CAM (CAM5) show a relatively larger thermodynamic bias. Li et al. (2020a) reported positive CAPE bias in CAM6 over the US resulting from surface moisture bias. The S06 patterns also suggest a similar behavior where BCC-CSM2-MR, CanESM5, MRI-ESM2.0, and MPI-ESM-1.2-HR provide more realistic representation, whereas models such as CMCC-CM2-SR5, two CNRM models, FGOALS-g3, MIROC6 have a relatively weaker shear magnitude.

The analysis of regional bias in convective parameters suggests significant variability across the three sub-domains over CONUS. We find that most models struggle to accurately simulate CAPE, CIN, and wind shear over the western CONUS, where SCS activity is relatively lower. However, models show an improved skill over the central CONUS region, where SCS activity is more frequent. Over the eastern CONUS sub-domain, CAPE is generally overestimated, whereas CIN and wind shear are mostly underestimated. Most models considerably underestimate S01, with poor representation of spatial variability. This likely impacts the representation of thunderstorm

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

environments associated with tornadoes (Markowski and Richardson 2014). We find that the biases in thresholded frequencies are much larger than the biases in mean patterns. Specifically, there is a substantial negative bias in thresholded frequency of summertime wind shear in most models. These biases highlight the importance of considering more thunderstorm relevant characteristics, and imply potential for appreciable biases. This finding is particularly significant, given that many previous studies investigating severe thunderstorm environments often emphasize the product of CAPE and S06 (or its variants). Such an approach might overlook potential biases in S06, impacting both its spatial distribution and magnitude.

The results from the present study provide a comprehensive evaluation of convective-storm parameters simulated by 12 CMIP6 models with respect to two independent reanalysis products. Although the reanalysis datasets suffer from biases (Taszarek et al. 2021c,a), they provide a temporally consistent way to validate the models. We believe that these results provide a valuable resource for selecting models that are applied either in statistical downscaling using future thunderstorm environments, or in dynamic downscaling. However, we do not explicitly highlight any model as the "best model" primarily because all of the models considered here exhibit biases in component parameters, meaning that selection must depend on the application. On this basis, we would argue that studies analyzing future thunderstorms over the US should opt for an ensemble approach to encompass the substantial degree of model-to-model variability in both thermodynamic and kinematic parameters.

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

Acknowledgments. This research was supported by Aon Inc. The authors acknowledge the feedback and comments of two anonymous reviewers during the review process.

Data availability statement. The 6-hourly CMIP6 datasets in zarr format is available on Google Cloud storage at `https://cloud.google.com/blog/products/data-analytics/new-climate-model-data-now-google-public-datasets`. The CMIP6 model data are accessible through `https://esgf-node.llnl.gov/search/cmip6`. The ERA5 and MERRA2 reanalysis products are available from `https://cds.climate.copernicus.eu` and `https://disc.gsfc.nasa.gov/datasets?project=MERRA-2` respectively. The python package *xcape* used for the computation of convective parameters is available at `https://zenodo.org/records/5270332`.

References

Allen, J. T., 2018: Climate change and severe thunderstorms. *Oxford Research Encyclopedia of Climate Science*.

Allen, J. T., D. J. Karoly, and K. J. Walsh, 2014a: Future Australian severe thunderstorm environments. Part II: The influence of a strongly warming climate on convective environments. *Journal of Climate*, **27 (10)**, 3848–3868.

Allen, J. T., D. J. Karoly, and K. J. Walsh, 2014b: Future australian severe thunderstorm environments. part ii: The influence of a strongly warming climate on convective environments. *Journal of Climate*, **27 (10)**, 3848–3868.

Ashley, W. S., A. M. Haberlie, and V. A. Gensini, 2023: The Future of Supercells in the United States. *Bulletin of the American Meteorological Society*, **104 (1)**, E1–E21.

Brooks, H. E., 2013: Severe thunderstorms and climate change. *Atmospheric research*, **123**, 129–138.

Brooks, H. E., J. W. Lee, and J. P. Craven, 2003: The spatial distribution of severe thunderstorm and tornado environments from global reanalysis data. *Atmospheric Research*, **67**, 73–94.

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

Chavas, D. R., and F. Li, 2022: Biases in CMIP6 Historical US Severe Convective Storm Environments Driven by Biases in Mean-State Near-Surface Moist Static Energy. *Geophysical Research Letters*, **49 (23)**, e2022GL098 527.

Cherchi, A., and Coauthors, 2019: Global mean climate and main patterns of variability in the CMCC-CM2 coupled model. *Journal of Advances in Modeling Earth Systems*, **11 (1)**, 185–209.

Craven, J. P., H. E. Brooks, J. A. Hart, and Coauthors, 2004: Baseline climatology of sounding derived parameters associated with deep, moist convection. *Natl. Wea. Dig*, **28 (1)**, 13–24.

Danabasoglu, G., and Coauthors, 2020: The community earth system model version 2 (CESM2). *Journal of Advances in Modeling Earth Systems*, **12 (2)**, e2019MS001 916.

Diffenbaugh, N. S., M. Scherer, and R. J. Trapp, 2013: Robust increases in severe thunderstorm environments in response to greenhouse forcing. *Proceedings of the National Academy of Sciences*, **110 (41)**, 16 361–16 366.

Doswell III, C. A., 2003: Societal impacts of severe thunderstorms and tornadoes: Lessons learned and implications for Europe. *Atmospheric Research*, **67**, 135–152.

Doswell III, C. A., and E. N. Rasmussen, 1994: The effect of neglecting the virtual temperature correction on CAPE calculations. *Weather and forecasting*, **9 (4)**, 625–629.

Durre, I., X. Yin, R. S. Vose, S. Applequist, and J. Arnfield, 2018: Enhancing the data coverage in the integrated global radiosonde archive. *Journal of Atmospheric and Oceanic Technology*, **35 (9)**, 1753–1770.

Emanuel, K., 2023: On the physics of high CAPE. *Journal of the Atmospheric Sciences*, **80 (11)**, 2669–2683.

Emanuel, K. A., and Coauthors, 1994: *Atmospheric convection*. Oxford University Press on Demand.

Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor, 2016: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, **9 (5)**, 1937–1958.

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

Gelaro, R., and Coauthors, 2017: The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). *Journal of climate*, **30 (14)**, 5419–5454.

Gensini, V. A., and T. L. Mote, 2014: Estimations of hazardous convective weather in the United States using dynamical downscaling. *Journal of Climate*, **27 (17)**, 6581–6589.

Gensini, V. A., and T. L. Mote, 2015: Downscaled estimates of late 21st century severe weather from CCSM3. *Climatic Change*, **129 (1)**, 307–321.

Haberlie, A. M., W. S. Ashley, C. M. Battisto, and V. A. Gensini, 2022: Thunderstorm activity under intermediate and extreme climate change scenarios. *Geophysical Research Letters*, **49 (14)**, e2022GL098 779.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, **146 (730)**, 1999–2049.

Hoeppe, P., 2016: Trends in weather related disasters–Consequences for insurers and society. *Weather and climate extremes*, **11**, 70–79.

Hoogewind, K. A., M. E. Baldwin, and R. J. Trapp, 2017: The impact of climate change on hazardous convective weather in the United States: Insight from high-resolution dynamical downscaling. *Journal of Climate*, **30 (24)**, 10 081–10 100.

Huang, J., 2018: A simple accurate formula for calculating saturation vapor pressure of water and ice. *Journal of Applied Meteorology and Climatology*, **57 (6)**, 1265–1272.

IPCC, 2023: Climate Change 2023: Synthesis Report. A Report of the Intergovernmental Panel on Climate Change. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland, 35–115 pp.

Kelley, M., and Coauthors, 2020: GISS-E2. 1: Configurations and climatology. *Journal of Advances in Modeling Earth Systems*, **12 (8)**, e2019MS002 025.

King, A. T., and A. D. Kennedy, 2019: North American supercell environments in atmospheric reanalyses and RUC-2. *Journal of Applied Meteorology and Climatology*, **58 (1)**, 71–92.

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

Lepore, C., R. Abernathey, N. Henderson, J. T. Allen, and M. K. Tippett, 2021a: Future global convective environments in CMIP6 models. *Earth's Future*, **9 (12)**, e2021EF002 277.

Lepore, C., J. T. Allen, and R. Abernathey, 2021b: XCAPE v0.1.4. URL https://doi.org/10.5281/zenodo.5270332.

Li, F., D. R. Chavas, K. A. Reed, and D. T. Dawson II, 2020a: Climatology of severe local storm environments and synoptic-scale features over North America in ERA5 reanalysis and CAM6 simulation. *Journal of Climate*, **33 (19)**, 8339–8365.

Li, L., and Coauthors, 2020b: The flexible global ocean-atmosphere-land system model grid-point version 3 (FGOALS-g3): description and evaluation. *Journal of Advances in Modeling Earth Systems*, **12 (9)**, e2019MS002 012.

Lin, Y., W. Dong, M. Zhang, Y. Xie, W. Xue, J. Huang, and Y. Luo, 2017: Causes of model dry and warm bias over central US and impact on climate projections. *Nature Communications*, **8 (1)**, 881.

Markowski, P. M., and Y. P. Richardson, 2014: The influence of environmental low-level shear and cold pools on tornadogenesis: Insights from idealized simulations. *Journal of the Atmospheric Sciences*, **71 (1)**, 243–275.

Marsh, P. T., H. E. Brooks, and D. J. Karoly, 2007: Assessment of the severe weather environment in North America simulated by a global climate model. *Atmospheric Science Letters*, **8 (4)**, 100–106.

Mauritsen, T., and Coauthors, 2019: Developments in the MPI-M Earth System Model version 1.2 (MPI-ESM1. 2) and its response to increasing CO2. *Journal of Advances in Modeling Earth Systems*, **11 (4)**, 998–1038.

Molina, M. J., and J. T. Allen, 2019: On the moisture origins of tornadic thunderstorms. *Journal of Climate*, **32 (14)**, 4321–4346.

Molina, M. J., and J. T. Allen, 2020: Regionally-stratified tornadoes: Moisture source physical reasoning and climate trends. *Weather and Climate Extremes*, **28**, 100 244.

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

Molina, M. J., J. T. Allen, and V. A. Gensini, 2018: The gulf of mexico and enso influence on subseasonal and seasonal conus winter tornado variability. *Journal of Applied Meteorology and Climatology*, **57 (10)**, 2439–2463.

O'Gorman, P., and C. J. Muller, 2010: How closely do changes in surface and column water vapor follow Clausius–Clapeyron scaling in climate change simulations? *Environmental Research Letters*, **5 (2)**, 025 207.

Pilguj, N., M. Taszarek, J. T. Allen, and K. A. Hoogewind, 2022: Are trends in convective parameters over the United States and Europe consistent between reanalyses and observations? *Journal of Climate*, **35 (12)**, 3605–3626.

Rasmussen, E. N., and D. O. Blanchard, 1998: A baseline climatology of sounding-derived supercell and tornado forecast parameters. *Weather and forecasting*, **13 (4)**, 1148–1164.

Rasmussen, K. L., A. F. Prein, R. M. Rasmussen, K. Ikeda, and C. Liu, 2020: Changes in the convective population and thermodynamic environments in convection-permitting regional climate simulations over the United States. *Climate Dynamics*, **55**, 383–408.

Seeley, J. T., and D. M. Romps, 2015: The effect of global warming on severe thunderstorms in the United States. *Journal of Climate*, **28 (6)**, 2443–2458.

Séférian, R., and Coauthors, 2019: Evaluation of CNRM Earth System Model, CNRM-ESM2-1: role of Earth system processes in present-day and future climate. *Journal of Advances in Modeling Earth Systems*, **11 (12)**, 4182–4227.

Seland, Ø., and Coauthors, 2020: Overview of the Norwegian Earth System Model (NorESM2) and key climate response of CMIP6 DECK, historical, and scenario simulations. *Geoscientific Model Development*, **13 (12)**, 6165–6200.

Skinner, C. B., and N. S. Diffenbaugh, 2013: The contribution of African easterly waves to monsoon precipitation in the CMIP3 ensemble. *Journal of Geophysical Research: Atmospheres*, **118 (9)**, 3590–3609.

Swart, N. C., and Coauthors, 2019: The Canadian earth system model version 5 (CanESM5. 0.3). *Geoscientific Model Development*, **12 (11)**, 4823–4873.

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

Taszarek, M., J. T. Allen, H. E. Brooks, N. Pilguj, and B. Czernecki, 2021a: Differing trends in United States and European severe thunderstorm environments in a warming climate. *Bulletin of the American Meteorological society*, **102 (2)**, E296–E322.

Taszarek, M., J. T. Allen, M. Marchio, and H. E. Brooks, 2021b: Global climatology and trends in convective environments from ERA5 and rawinsonde data. *NPJ climate and atmospheric science*, **4 (1)**, 35.

Taszarek, M., N. Pilguj, J. T. Allen, V. Gensini, H. E. Brooks, and P. Szuster, 2021c: Comparison of convective parameters derived from ERA5 and MERRA-2 with rawinsonde data over Europe and North America. *Journal of Climate*, **34 (8)**, 3211–3237.

Tatebe, H., and Coauthors, 2019: Description and basic evaluation of simulated mean state, internal variability, and climate sensitivity in MIROC6. *Geoscientific Model Development*, **12 (7)**, 2727–2765.

Trapp, R. J., 2013: *Mesoscale-convective processes in the atmosphere*. Cambridge University Press.

Trapp, R. J., N. S. Diffenbaugh, H. E. Brooks, M. E. Baldwin, E. D. Robinson, and J. S. Pal, 2007: Changes in severe thunderstorm environment frequency during the 21st century caused by anthropogenically enhanced global radiative forcing. *Proceedings of the National Academy of Sciences*, **104 (50)**, 19 719–19 723.

Trapp, R. J., K. A. Hoogewind, and S. Lasher-Trapp, 2019: Future changes in hail occurrence in the United States determined through convection-permitting dynamical downscaling. *Journal of Climate*, **32 (17)**, 5493–5509.

Trapp, R. J., E. D. Robinson, M. E. Baldwin, N. S. Diffenbaugh, and B. R. Schwedler, 2011: Regional climate of hazardous convective weather through high-resolution dynamical downscaling. *Climate dynamics*, **37**, 677–688.

Tuckman, P., V. Agard, and K. Emanuel, 2023: Evolution of convective energy and inhibition before instances of large cape. *Monthly Weather Review*, **151 (1)**, 321–338.

Voldoire, A., and Coauthors, 2019: Evaluation of CMIP6 deck experiments with CNRM-CM6-1. *Journal of Advances in Modeling Earth Systems*, **11 (7)**, 2177–2213.

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.

Wang, Z., J. A. Franke, Z. Luo, and E. J. Moyer, 2021: Reanalyses and a high-resolution model fail to capture the "high tail" of cape distributions. *Journal of Climate*, **34 (21)**, 8699–8715.

Wu, T., and Coauthors, 2019: The Beijing Climate Center Climate System Model (BCC-CSM): the main progress from CMIP5 to CMIP6. *Geoscientific Model Development*, **12 (4)**, 1573–1600.

Yukimoto, S., and Coauthors, 2019: The Meteorological Research Institute Earth System Model version 2.0, MRI-ESM2. 0: Description and basic evaluation of the physical component. *Journal of the Meteorological Society of Japan. Ser. II*, **97 (5)**, 931–965.

Zhang, G. J., 2009: Effects of entrainment on convective available potential energy and closure assumptions in convection parameterization. *Journal of Geophysical Research: Atmospheres*, **114 (D7)**.

Brought to you by University of Maryland, McKeldin Library | Unauthenticated | Downloaded 01/22/25 10:16 PM UTC

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-24-0165.1.