Predictability of Severe Convective Storm Environments in Global Ensemble Forecast System, Version 12, Reforecasts

ANDREW H. BERRINGTON,[a,b,c] KIMBERLY A. HOOGEWIND,[a,b] ADAM J. CLARK,[b,c] AND MATEUSZ TASZAREK[d]

[a] *Cooperative Institute for Severe and High Impact Weather Research and Operations, University of Oklahoma, Norman, Oklahoma*
[b] *NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma*
[c] *School of Meteorology, University of Oklahoma, Norman, Oklahoma*
[d] *Department of Climatology, Adam Mickiewicz University, Poznań, Poland*

(Manuscript received 16 January 2024, in final form 8 August 2024, accepted 23 August 2024)

ABSTRACT: Prediction of severe convective storms at time scales of 2–4 weeks is of interest to forecasters and stakeholders due to their impacts on life and property. Prediction of severe convective storms on this time scale is challenging, since the large-scale weather patterns that drive this activity begin to lose dynamic predictability beyond week 1. Previous work related to severe convective storms on the subseasonal time scale has mostly focused on observed relationships with teleconnections. The skill of numerical weather prediction forecasts of convective-related variables has been comparatively less explored. In this study over the United States, a forecast evaluation of variables relevant in the prediction of severe convective storms is conducted using Global Ensemble Forecast System, version 12, reforecasts at lead times of up to 4 weeks. We find that kinematic and thermodynamic fields are predicted with skill out to week 3 in some cases, while composite parameters struggle to achieve meaningful skill into week 2. Additionally, using a novel method of weekly summations of daily maximum composite parameters, we suggest that the aggregation of certain variables may assist in providing additional predictability beyond week 1. These results should serve as a reference for forecast skill for the relevant fields and help inform the development of convective forecasting tools at time scales beyond current operational products.

SIGNIFICANCE STATEMENT: Prediction of severe weather beyond 1 week in advance is of interest to stakeholders given their impacts on life and property. In this study, we evaluate 20 years of forecast data generated by a numerical model ensemble to determine whether variables relevant in severe weather forecasts can be predicted in weeks 2–4. The variables with the best skill measures generally represent large-scale weather patterns that are more predictable on longer time scales, although some manipulation of other severe weather parameters yielded additional results. We suggest that the results found in this study can inform future work that assesses the predictability of severe weather in weeks 2–4 using more complex methods such as machine learning.

KEYWORDS: Severe storms; Forecast verification/skill; Hindcasts

1. Introduction

Severe convective storms (SCSs) and their hazards including tornadoes, damaging convective wind gusts, and large hail represent a significant threat to life and property across the contiguous United States (CONUS), with an estimated $10 billion in insured losses each year (Smith and Katz 2013). Given these societal impacts, skillful subseasonal (defined here as weeks 2–4) prediction of SCSs could benefit emergency managers, the agricultural sector, the reinsurance industry, and other stakeholders (Craig et al. 2021). However, subseasonal prediction of SCSs is challenging, as variables (shear and instability) relevant in SCS forecasts—such as those considered in Storm Prediction Center (SPC) convective outlooks (Hitchens and Brooks 2014; Herman et al. 2018) that extend to day 8—tend to lose meaningful predictive skill at the subseasonal time scale. At a more basic level, the intrinsic predictability of an individual SCS is confined to much shorter time scales than a week (Lorenz 1969; Melhauser and Zhang 2012; Markowski 2020).

The current body of the literature assessing SCSs on the subseasonal time scale primarily focuses on statistical links between observed climate teleconnection indices and SCS activity. The Madden–Julian oscillation (MJO) (Madden and Julian 1972; Wheeler and Hendon 2004; Zhang 2005; Kiladis et al. 2014; Tseng et al. 2018) has been identified as a potential source of predictability for SCS activity during boreal spring, given the role of its extratropical response in modulating the North Pacific jet (NPJ) and resulting downstream impacts over North America (Tseng et al. 2018; Winters 2021). As part of support for the Extended-Range Tornado Activity Forecast (ERTAF) project (Gensini et al. 2020) and using the MJO and state of global atmospheric angular momentum (Gensini and Marinaro 2016; Gensini and Allen 2018; Moore et al. 2018; Moore and McGuire 2020), a successful subseasonal prediction of an extended episode of SCSs in the second half of May 2019 was made (Gensini et al. 2019). Further works by Barrett and Gensini (2013), Barrett and Henley (2015), Baggett et al. (2018), and Miller et al. (2022) have elaborated on the influence of the MJO on tornado and hail activity across the CONUS, with varying results that are dependent on the choice of MJO indices and measure of SCS activity used (Tippett 2018).

Corresponding author: Andrew H. Berrington, andrew.berrington@ou.edu

DOI: 10.1175/WAF-D-24-0004.1

© 2024 American Meteorological Society. This published article is licensed under the terms of the default AMS reuse license. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 07:02 PM UTC

While climate oscillations and teleconnections have been examined for their influence on SCSs, studies to determine predictability within operational numerical weather prediction (NWP) models have been comparatively fewer. Gensini and Tippett (2019) used National Centers for Environmental Prediction (NCEP) day 1–15 Global Ensemble Forecast System, version 10 (GEFSv10), data and found that daily tornado and hail activity relative to thresholds could be forecast with skill compared to climatology out to day 9 and day 12, respectively. However, only two severe weather seasons were captured in their analysis and the only variable used as a predictor was the daily maximum supercell composite parameter (SCP). Additionally, they did not evaluate the intrinsic skill of the SCP within the model. Wang et al. (2021) built a statistical–dynamical hybrid model using the SCP from GEFS hindcasts and known relationships with SCS frequency. They found that low skill was obtained during week 2 by applying their model to higher-resolution data, while skill improved when applying area averaging to their environmental fields. Additionally, applying a singular value decomposition to the covariance between weekly storm reports and SCP improved forecast skill. However, it was only tested for a short period during 2019 and the inclusion of damaging winds in their severe weather reports introduces further uncertainties due to substantial changes over time in reporting methods.

The study of Lepore et al. (2018) using Climate Forecast System, version 2 (CFSv2), forecasts (Saha et al. 2014) found that much of the monthly skill for convective available potential energy (CAPE), convective precipitation, and storm-relative helicity (SRH) was concentrated within the first 2 weeks of the extended forecast, which corresponds to when dynamical predictability within global models still exists. The CFSv2 was also used in the seasonal prediction of tornadoes using large-scale variability described in Lee et al. (2021) and in the visualization of midrange SCS environments presented by Carbin et al. (2016), although limitations of the model itself create difficulties in its use for predicting extreme events on these time scales. Carbin et al. (2016) documented that run-to-run differences were substantial beyond week 1 for their SCP aggregations. Notably, Lee et al. (2021) did not evaluate the intrinsic skill or covariability of the parameters (low-level vertical wind shear and instability) chosen from the CFSv2 in their forecasting method and used a different background climatology from a separate reanalysis to create anomalies. Broadly, the lack of skill evaluation of raw output variables is a common theme in previous studies attempting to use proxies to forecast severe weather at longer leads.

Given that planetary-scale climate teleconnections are unable to explain a large portion of variability in SCS activity (Moore et al. 2018), promising work in recent years has also been conducted surrounding synoptic weather regimes (WRs) favorable for SCSs. By using *K*-means clustering, Miller et al. (2020) yielded five dominant WRs in reanalysis data during the month of May in 500-hPa geopotential height across the CONUS. A large percentage (75%) of SCS outbreak days within their study occurred during persistent WRs where a western mid- and upper-level thermal trough was present and where the magnitude and spatial coverage of instability- and shear-related variables increase relative to climatology. Subsequent analysis utilizing a hybrid prediction model blending European Centre for Medium-Range Weather Forecasts (ECMWF) data and information about the WRs for weekly SCS activity yielded skillful forecasts relative to climatology out to week 3. Despite these results, Miller et al. (2020) only assessed 1 month of the year and also highlighted that they did not evaluate the predictive skill of the ECMWF, which would impact the skill of this approach. Subsequently, a promising approach of using machine learning techniques has been implemented in short- and medium-range SCS forecasts (Hill et al. 2020, 2023), which has yielded results comparable to SPC outlooks for these time scales. However, such machine learning methods are also limited by the skill of the model forecast parameters that are used as inputs.

In line with improving knowledge of extended range NWP for a variety of weather phenomena, reforecast datasets have become a powerful tool for examining longer climatologies of model prediction (Hamill et al. 2006). NCEP's GEFS reforecasts (Hamill et al. 2013)—hereby referred to as GEFSR—have been used for a variety of applications including precipitation forecasting (Hamill et al. 2008; Baxter et al. 2014; Herman and Schumacher 2016; Guan et al. 2022), evapotranspiration and drought (Shah and Mishra 2015, 2016; Mo and Plettenmaier 2020; Talib et al. 2021), and even aviation hazards (Verlinden and Bright 2017). However, SCS prediction using GEFSR or other reforecast data has not been extensively researched, despite climatologies of model reanalysis being established from the observational perspective (Taszarek et al. 2020; Li et al. 2020). With the release of the updated GEFSR, version 12 (GEFSRv12) as documented in Zhou et al. (2022), evaluating forecast climatology for a state-of-the-art NWP ensemble could provide additional clues into expanding the predictability window for SCSs beyond current operational capabilities.

The GEFSv12 has shown documented improvements in both its reforecast (Guan et al. 2022) and reanalysis (Hamill et al. 2022) products, including lessening of its temperature biases near the surface, although snow cover still presents a challenge. Guan et al. (2022) additionally documented that Northern Hemisphere 500-hPa heights exhibited better skill via anomaly correlations during weeks 1–2 than previous iterations of the GEFS, while additionally improving precipitation and MJO forecasts. However, most variables relevant for SCS forecasting were not evaluated in their work. A notable recent application of the GEFSv12 toward SCS forecasting was conducted by Miller and Gensini (2023), who found relationships between the skill of ensemble forecasts in the extended range and global sea surface temperature patterns, wave activity flux, blocking regimes, and MJO activity. Low-skill forecasts for SCS days were characterized by synoptic features in 500-hPa heights propagating too quickly compared to observations. Atmospheric blocking regimes over the North Pacific contributed to these errors. Overall, Miller and Gensini (2023) provided a promising first look at "predicting skill" in the GEFSv12 for SCS events, although the intrinsic skill of variables outside of 500-hPa heights was not explored.

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 07:02 PM UTC

As a general assessment, the deterministic limit for midlatitude weather predictability is roughly 10 days (Zhang et al. 2019; Miller and Gensini 2023) and is certainly shorter for more localized phenomena such as SCSs. However, relaxing time and space constraints on forecast verification may yield additional skill at longer lead times (Buizza and Leutbecher 2015). Probabilistic verification methods may additionally be more useful for end users of subseasonal forecasts (Manrique-Suñén et al. 2020).

With the above considerations in mind, this study seeks to establish baseline skill (Goutham et al. 2022) of variables relevant to subseasonal SCS forecasting within GEFSRv12. Such explicit skill evaluation has been mostly absent from the previous work (Gensini and Tippett 2019; Lee et al. 2021) and can help inform the choice of predictors when considering proxies for SCS activity. We will relax time and space constraints for forecast verification as we do not intend to forecast for subdaily time frames at subseasonal time scales. This is primarily due to potential timing differences of pertinent features between ensemble members and observations due to error growth in subseasonal forecasts. Both probabilistic and deterministic methods are employed to assess the quality of the reforecasts. While regional focuses are possible among climatology, we emphasize that the primary goal of this study is to address the above questions on a CONUS-wide basis first.

2. Data and methods

a. *Reforecast and observational data*

Twenty years of GEFSRv12 data from 2000 to 2019 are used in this study, with a mixed temporal resolution of 3 hours for the first 10 days and 6 h for the remainder of the 35-day forecasts. The study domain is centered on the CONUS and spans 20°–50°N latitude and 140°–55°W longitude. For the purposes of this study, only 6-hourly increments are factored into mean calculations, to maintain consistency between reforecasts and reanalysis. Reforecasts out to 35 days are available once per week, initialized at 0000 UTC, during this time span with 11 forecast members, while reforecasts out to 16 days are available daily with only five members. For the purposes of this work, only the weekly 35-day forecasts are used to evaluate forecast skill during weeks 2–4. For the 20 years of data, this corresponds to 1028 total forecasts after removing 15 forecast cycles that contained missing data. All near-surface and pressure level data up to 700 hPa are horizontally regridded from its 0.25° latitude–longitude grid spacing to 0.5° using bilinear interpolation via the xESMF package in Python (Zhuang et al. 2022). Data above 700 hPa are kept at its available 0.5° grid spacing (Guan et al. 2022). Additionally, daily means for all variables are constructed from 1200 to 1200 UTC from the original 6-hourly data, in order to match the report periods for SPC local storm reports (LSRs). Observational atmospheric fields are derived from GEFSv12 reanalysis over the same 20-yr period. The reanalysis data are regridded and temporally averaged similarly to the reforecast data, in order to provide direct comparison in subsequent verification metrics.

While assessing daily means and maxima of quantities relevant in SCS forecasting is helpful, it may not be sufficient to do so at the subseasonal time scale, where the prediction of key features in an ensemble mean may become difficult due to timing differences (Gensini and Tippett 2019; Manrique-Suñén et al. 2020). With the objective of reducing the impact of timing differences for higher frequency variability at the daily scale (Li and Stechmann 2018; Guan et al. 2022), we compute 7-day forward-running means from the reanalysis and reforecast variables and compare them to the daily means to assess how expanding the temporal window impacts skill. However, these weekly averages come at the sacrifice of forecast resolution and sharpness, potentially smoothing over important features. For the purposes of both deterministic and probabilistic verification scores, we select a set of 16 relevant quantities comprising three kinematic/mass-related variables, four thermodynamic variables, five individual convective parameters, and four composite convective parameters that are listed below. The shorthand names for these variables in parentheses are used hereafter in the manuscript text.

- Kinematic/mass variables: 500-hPa geopotential height (z500; gpm), 250-hPa zonal wind (u250; m s^{-1}), and 850-hPa meridional wind (v850; m s^{-1}).
- Thermodynamic variables: 850-hPa specific humidity (q850; g kg^{-1}), 2-m temperature (t2m; K), 2-m specific humidity (q2m; g kg^{-1}), and 700-hPa temperature (t700; K).
- Individual convective parameters: surface-based CAPE (SBCAPE; J kg^{-1}), 0–3-km AGL SRH (m^2 s^{-2}), 0–6-km bulk wind shear (BS06; m s^{-1}), 0–1-km BS (BS01; m s^{-1}), and surface-based convective inhibition (SBCIN; J kg^{-1}).
- Composite convective parameters: Fixed-layer SCP (SCP-fixed), fixed-layer significant tornado parameter (STP-fixed), the SHERB parameter (Sherburn and Parker 2014) using effective BS (SHERBE), and the product of 0–6-km BS and most unstable CAPE (CAPE–SHEAR).

As convective parameters—and particularly composite severe parameters such as STP (Thompson et al. 2003; Grams et al. 2012)—frequently register zero values due to their inherent climatologies (e.g., instability during the wintertime) or are muted due to their strong diurnal cycles (e.g., instability with daytime heating) when generating a daily or longer mean value, we instead use daily maxima or minima in the 1200 to 1200 UTC window for these quantities. Daily maxima are used for SBCAPE, SCP-fixed, STP-fixed, SHERBE, and CAPE–SHEAR, while daily minima are used for SBCIN. However, this alone may not be sufficient to capture variability over longer time scales for composite SCS parameters (Gensini and De Guenni 2019). Therefore, weekly sums of daily maximum SCP-fixed, STP-fixed, SHERBE, and CAPE–SHEAR analogous to the Craven–Brooks significant severe parameter (Craven and Brooks 2004) are computed to examine aggregated SCS indices in the GEFS reforecast data against reanalysis. Equations for the composite convective parameters used in this study are given in the appendix.

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 07:02 PM UTC

b. Forecast evaluation metrics

A variety of forecast verification techniques are used in this study to gauge the skill of the GEFSRv12 in identifying patterns in variables relevant for SCS forecasting. Deterministic measures include anomaly correlation coefficients (ACCs), root-mean-square errors (RMSEs), and ensemble consistency diagrams (Eckel and Mass 2005; Clark et al. 2010) to assess the relationship of ensemble variance and mean-square error. When computing seasonal delineations, we use reforecasts initialized within the given season, add verification dates that are within the given season, and remove verification dates that are in the following season. For example, March–May (MAM) reforecast verification dates extend into June at the end of the season and the forecast initialization dates at the beginning of the season are in February. In this case, the former is removed and the latter is added to computations of skill metrics, when applicable. Anomalies for all verification metrics are computed by subtracting the centered 31-day running mean (the mean on May 1 would be from 15 April to 15 May) climatology derived from reanalysis from the daily reanalysis or reforecast data. This is performed prior to any further temporal averaging. Subsequently, the daily anomalies are either kept as is for daily verification metrics or smoothed using the forward-running window mentioned above for the weekly verification metrics.

While deterministic skill scores can give a general overview as to the performance of a model, they do not necessarily capture the likelihood of a "skilled" forecast verification. Particularly at longer lead times when ensemble spread increases, timing differences of key features may be significant and result in time-ensemble means trending toward climatology. Probabilistic skill scores offer a more comprehensive assessment of performance for ensemble forecasts (Weigel et al. 2007; Wilks 2011; Manrique-Suñén et al. 2020). Probabilistic metrics used in this study include rank histograms (Hamill 2001), receiver operator characteristic (ROC) curves (Mason 1982), reliability diagrams (Hamill 1997; Bröcker and Smith 2007), and Brier skill scores or ranked probability skill scores (BSSs or RPSSs) measured against model climatology (Epstein 1969; Weigel et al. 2007; Wilks 2011; Gensini and Tippett 2019; Manrique-Suñén et al. 2020). In this study, rank histograms are composed of 12 ranks, grouping 11 ensemble members and one observation from reanalysis. The ROC curves represent the relationship between the true-positive rate (TPR) and false-positive rate (FPR) of the ensemble. Area under the ROC curve (AUC) is calculated using the trapezoidal integration method. Meanwhile, the RPSS quantifies how well an ensemble is performing at predicting probability distribution, usually split into categorical bins, against a reference (usually climatology) forecast. The BSS is a special case of the RPSS where there is a binary set of outcomes (yes or no). Finally, the reliability diagram is simply the relationship between predicted probabilities of a given outcome (e.g., SCP exceeding a given threshold) and the observed true frequency of the outcome.

For the necessity of a reference (climatological) model skill in probabilistic skill scores, Manrique-Suñén et al. (2020) emphasized the importance of carefully choosing the lead-dependent climatology computation for subseasonal forecasts. Additional consideration of the underlying climatology is warranted (Hamill and Juras 2006), especially given the difference in the temporal scale between subseasonal and SCS phenomena. We opt to use a running-window monthly climatology to increase the sample size, similar to the third approach in Manrique-Suñén et al. (2020). Therefore, given the weekly temporal spacing between reforecasts, the targeted reforecast is chosen in addition to two reforecast initializations on either side of the targeted reforecast to compute the climatology. With 11 ensemble members and 20 years of reforecasts, this increases the sample size to 5 forecasts $\times$ 11 ensemble members $\times$ 20 years = 1100 at each lead time and grid point for the reforecast climatology. Using a similar rolling window, the observational reference yields a sample size of 5 observations $\times$ 20 years = 100 at each grid point. Both the actual forecast and lead-dependent climatology are then compared to the corresponding reanalysis and ranked probability (or Brier) scores are calculated, followed by the appropriate probabilistic skill score. All significance testing when applied is conducted at the 95% confidence level using bootstrapping with 5000 repetitions and independently at each grid point, unless otherwise indicated. Random selections in the bootstrap tests are not selected from specific time periods unless seasonal metrics are tested. Equations for select forecast verification metrics are presented in the appendix.

3. Results

a. Deterministic skill evaluation

Prior to conducting analysis on probabilistic forecasts using the ensemble members, it is beneficial to establish baseline climatological skill scores of the ensemble mean via deterministic measures. Two of the most popular deterministic metrics for forecast evaluation are the ACC and RMSE of the forecast compared to observations (Murphy and Epstein 1989; Joliffe and Stephenson 2012). Climatological ensemble mean anomaly correlation coefficients for the 16 selected variables relative to the reanalysis anomalies are presented in Fig. 1 across the CONUS domain and all seasons. These anomaly correlation coefficients are calculated for daily and 7-day (weekly) rolling means.

In Fig. 1a, the daily ensemble mean z500 anomaly has average ACC above the synoptic skill threshold (0.6) extending to roughly 9 days and climatological skill extending to 11 days (Guan et al. 2022; Miller and Gensini 2023), while weekly ensemble mean z500 extends the threshold for climatological skill (0.5) by around 1 day. This extension of forecast skill by temporally averaging fields has been documented in previous work (van Straaten et al. 2020) and extends to the remainder of the variables in Fig. 1. Shorter windows of skill are generally depicted for fields that are within the lower troposphere, particularly moisture-related variables. Various issues including boundary layer parameterization (Cohen et al. 2017; Hu et al. 2022) and timing of frontal passages may play roles in limiting predictability in the lower troposphere. Guan et al. (2022)

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 07:02 PM UTC

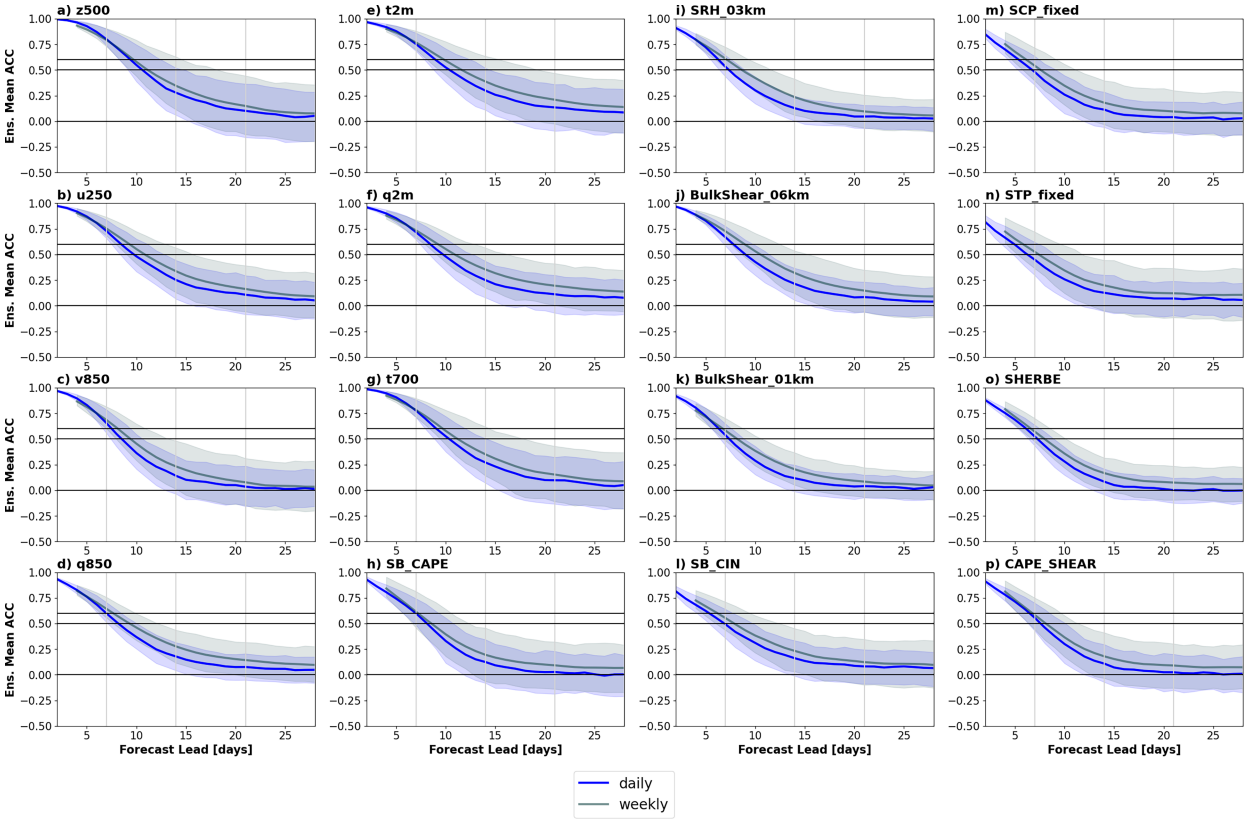

FIG. 1. Twenty-year climatological ensemble mean ACC for (a) z500, (b) u250, (c) v850, (d) q850, (e) t2m, (f) q2m, (g) t700, (h) SBCAPE, (i) SRH, (j) BS06, (k) BS01, (l) SBCIN, (m) SCP-fixed, (n) STP-fixed, (o) SHERBE, and (p) CAPE–SHEAR. The solid lines represent the mean, while the shaded regions represent the range between the 25th and 75th percentile ACC at each forecast lead. The blue lines represent the daily means (or maxima/minima), and the gray lines represent the 7-day centered rolling means (of daily means or maxima/minima). The horizontal solid black lines represent the thresholds for synoptic skill (0.6), climatological skill (0.5), and zero skill (0).

documented warm t2m biases during the warm season and cool t2m biases in the cool season—although substantially reduced from previous iterations of the GEFS—in their study that evaluated GEFSR, which may subsequently affect other thermodynamic fields such as SBCAPE.

As expected, given compounding errors with fields that incorporate multiple variables, both circulation and thermodynamic variables alone have higher ACC on average extending to longer lead times than composite SCS parameters. Noncomposite severe parameters show varying results, with BS06 and BS01 (Figs. 1j,k) mimicking the comparison between mid- and lower-tropospheric circulation variables, which is expected given the influence of the governing circulation patterns on vertical wind shear magnitude. Composite or vertically integrated SCS parameters show negligible skill beyond week 1 on average (Figs. 1h,i,l,m–p), in agreement with previous SCS studies (Gensini and Tippett 2019; Wang et al. 2021). In particular, both temporal averages of SBCIN decrease below synoptic-level skill using ACC before week 2. Weekly averaging extends climatological-level mean skill for SRH and SBCIN into week 2. Even prior to further analysis, this casts some degree of doubt in the fidelity of GEFSR in representing the likelihood of SCSs in the extended range, as convective inhibition can often make or break a potential SCS day. Composite parameters such as STP-fixed (Fig. 1n) that contain convective inhibition as a factor are likely to struggle when evaluated through deterministic metrics. Other quantities including t700 (Fig. 1g) that may be relevant for inhibition strength have a higher ACC compared to SBCIN. However, composite parameters such as STP-fixed (Fig. 1n) that contain convective inhibition as a factor are likely to struggle when evaluated through deterministic metrics.

While the bulk statistical means of the ACC provide a reasonable expectation for the skill of the ensemble mean, the 25th and 75th percentiles are also plotted to give bearish and bullish estimates. The latter can be considered comprising "forecasts of opportunity" (Gensini et al. 2019; Mariotti et al. 2020; Miller and Gensini 2023) where higher skill than average is obtained, although low spread must also equate with low error, which is evaluated in Fig. 3. For several variables including z500 (Fig. 1a) and t2m (Fig. 1e), this 75th percentile ACC extends climatological-level skill into week 3 for weekly means. However, even when considering the 75th percentile, composite convective parameters struggle to yield meaningful

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 07:02 PM UTC

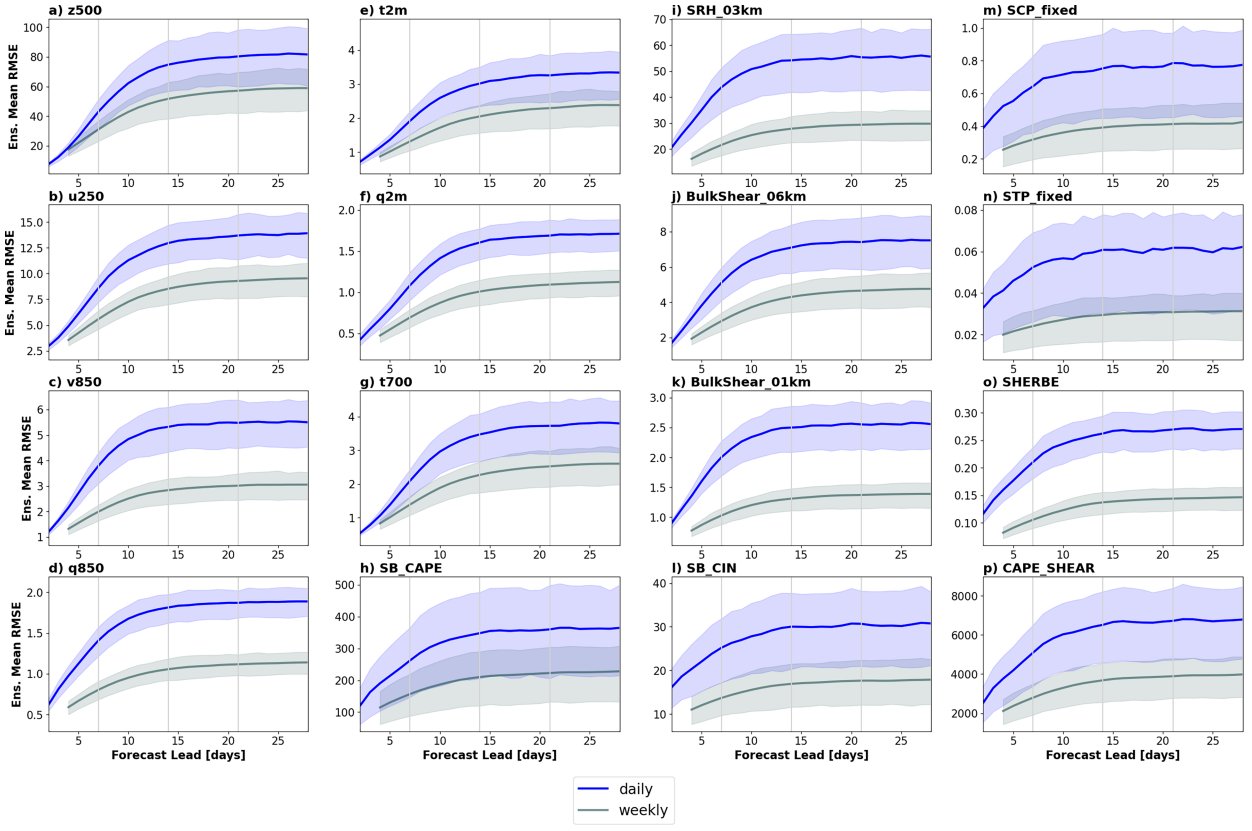

FIG. 2. As in Fig. 1, but for climatological RMSE.

forecasts beyond the middle of week 2 via ACC. Additionally, computing rolling averages of already smoothed daily maxes and means removes a considerable amount of information for parameters designed for forecasting rare events such as STP (van Straaten et al. 2020).

Figure 2 depicts the climatological RMSE for the 16 selected variables. The benefits of temporal generalization for deterministic verification are evident in several fields—particularly q850 (Fig. 2d)—where there is little-to-no overlap between the interquartile ranges for each temporal window. As expected, the saturation of errors also occurs at a longer lead for weekly means. However, for vertically integrated fields—such as SBCAPE and SBCIN (Figs. 2h,l)—along with composite parameters, this delineation is less clear. The SHERBE parameter (Fig. 2o) seems to yield better results for ACC and RMSE when a weekly mean is applied, which makes sense given its product consisting solely of 0–3-km and 700–500-hPa lapse rates and 0–3-km bulk wind shear, rather than encompassing vertically integrated quantities such as CAPE and CIN. Climatological RMSE thresholds are not shown in Fig. 2 in the interest of highlighting the differences between temporal averaging.

While not shown for brevity, seasonal cycles for both deterministic metrics generally follow those suggested by work in different regions (Büeler et al. 2021), where forecast skill is maximized during boreal winter and minimized during summer and early fall as the synoptic forcing from the midlatitude jet stream that provides additional predictability (Gensini et al. 2020; Winters 2021) weakens over the CONUS. During spring, mean and 75th percentile daily z500 ACC drops below key thresholds by 9–10-day and 13–14-day lead, respectively. This supports the exploratory work of Miller et al. (2020) using z500 WRs to predict SCSs beyond week 1 and the identification of forecasts of opportunity in Miller and Gensini (2023). Weekly averaging increases these by 1–2 days in the mean and 2–3 days at the 75th percentile.

In an 11-member ensemble, it is expected that underdispersion may be an issue, especially at shorter lead times. Ensemble consistency diagrams (Wang and Bishop 2003; Eckel and Mass 2005; Clark et al. 2010) describe the spread–skill relationship and dispersion of ensemble forecasts. For a well-dispersed ensemble, the relationship between ensemble variance and mean-square error should be close to 1:1. In Fig. 3, ensemble consistency is plotted as a function of lead time for daily mean variables. For kinematic and circulation fields, ensemble spread becomes a better predictor of mean error with time and r^2 values generally range from 0.3 to 0.6, indicating low to moderate variability explained. Temperature fields show similar temporal trends, but moisture fields suggest the opposite, particularly q2m (Fig. 3f). Forecasts in week 1 and week 2 also tend to be underdispersive, with the scatter skewed left of perfect consistency in a vertically oriented

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 07:02 PM UTC

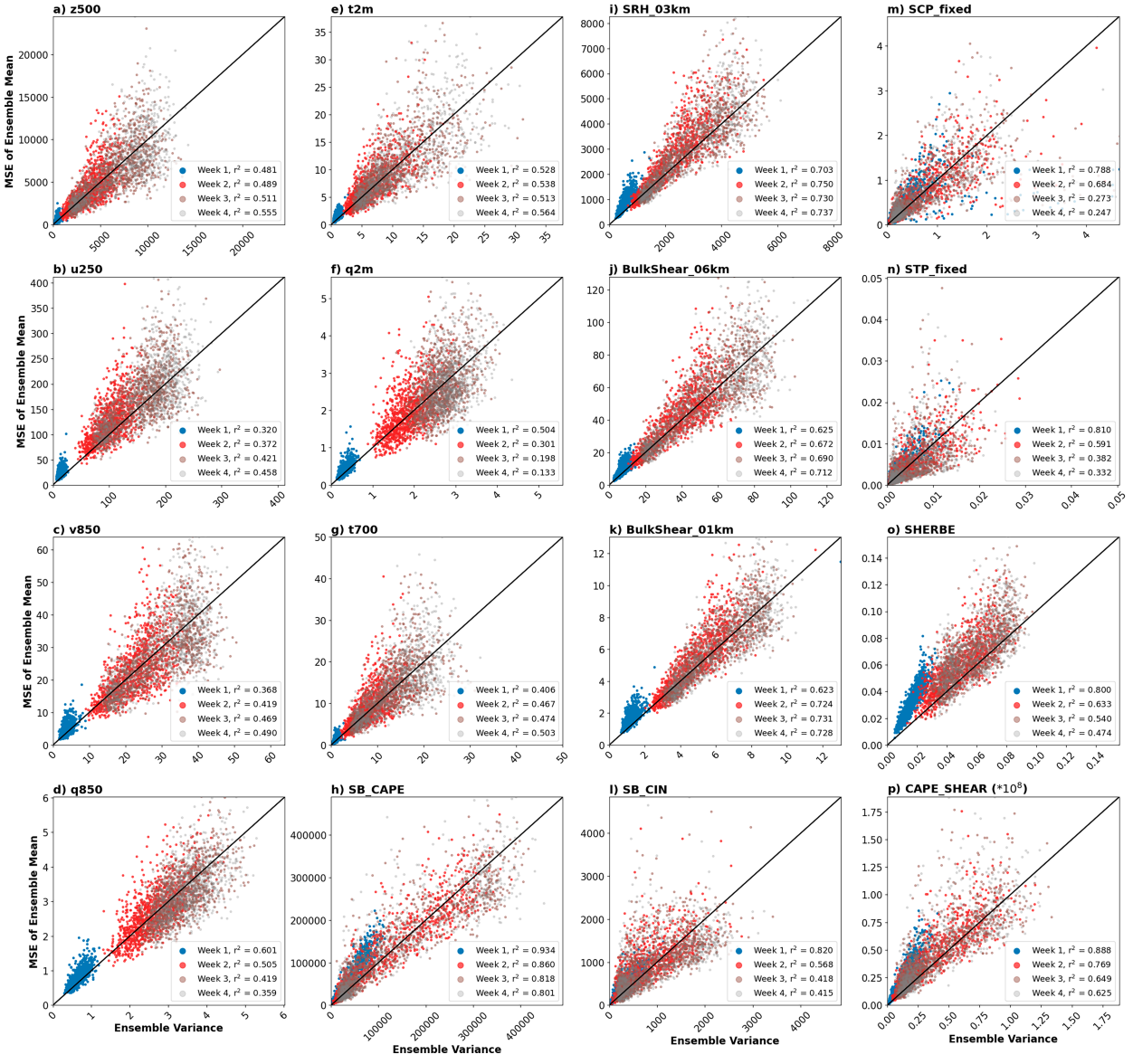

FIG. 3. Variables are as in Figs. 1 and 2, but for the relationship of ensemble variance and mean-square error calculated from daily means or maxima/minima. The scatter color indicates the week of the forecast lead, and the diagonal black line indicates a 1:1 relationship. Coefficients of determination r^2 for each week are indicated in the subplot legends.

plume (Clark et al. 2010). The behavior for SBCAPE (Fig. 3h) is inconsistent between forecast weeks, where overly dispersed forecasts in week 1 shift toward underdispersion by weeks 3 and 4. Meanwhile, SCP-fixed and STP-fixed (Figs. 3m,n) forecasts are generally overdispersed, especially during week 2. Similar results are found for the other temporal averages tested (not shown).

While the deterministic measures above suggest that kinematic, mass, and thermodynamic fields important in forecasting SCSs can be represented with fidelity out to at least week 2 in the GEFSRv12, they do not directly describe biases that may impact representation. Moreover, probabilistic measures capture generalized outcomes such as tercile verification, which may give a better indication of skill especially at longer lead times (Manrique-Suñén et al. 2020)—when timing differences affect deterministic metrics.

b. *Probabilistic skill evaluation*

As a complementary piece to ensemble consistency, rank histograms (Hamill 2001; Clark et al. 2010) describe the dispersion and biases of the ensemble compared to observations. For the purposes of this study, rank histograms for week 2 are evaluated, as the postweek 1 deterministic metrics yield the more promising results during this forecast lead as opposed to weeks 3 and 4. Weeks 3 and 4 exhibit very similar, yet slightly

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 07:02 PM UTC

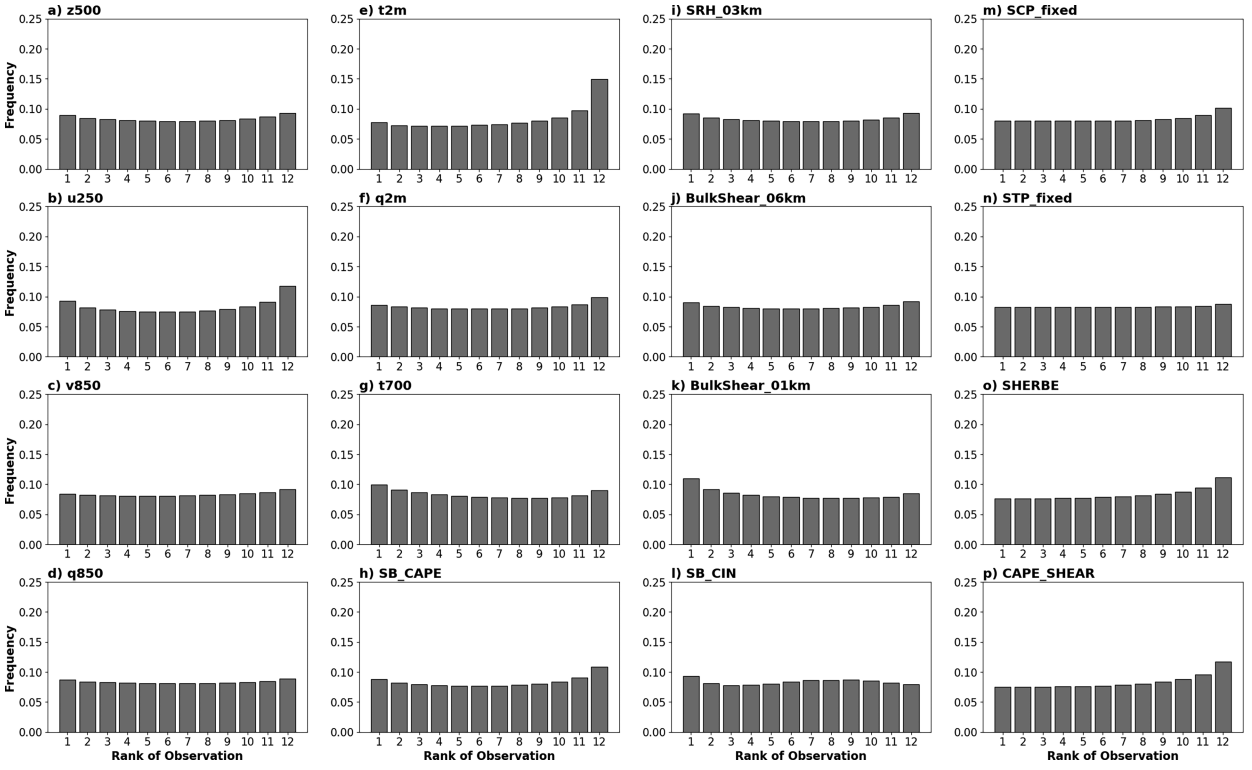

FIG. 4. Rank histograms for daily means of the same variables from Figs. 1 to 3 for all seasons during forecast week 2.

muted, biases as those found for week 2 (not shown) and overall ensemble dispersion is generally greater, as expected. Rank histograms for daily mean variables at week 2 for all seasons are presented in Fig. 4.

Examining kinematic fields at week 2 lead, the GEFSRv12 shows a reasonable amount of dispersion among ensemble members, with relatively flat rank histograms. There is a tendency toward underdispersion with upper-tropospheric fields such as u250 (Fig. 4b). Systematic biases are generally minimal, aside from a slight negative bias in u250. For thermodynamic fields, while forecasts for moisture are well dispersed and biases are weak, there is a cool bias evident in t2m (Fig. 4e) and a slight warm bias evident in t700 (Fig. 4g). The former supports the findings of Guan et al. (2022) and may impact the vertical profile and representation of parameters such as CAPE, CIN, and their derivatives. SBCAPE (Fig. 4h) does indeed have a modest low bias at week 2, while SBCIN (Fig. 4l) shows relatively little systematic bias in its rankings. SRH (Fig. 4i), BS06 (Fig. 4j), SCP-fixed (Fig. 4m), and STP-fixed (Fig. 4n) are generally well dispersed with minimal biases, while CAPE–SHEAR (Fig. 4p) and SHERBE (Fig. 4o) have slight low biases. BS01 (Fig. 4k) has a slight high bias. The disconnect between the magnitude of temperature biases and SCS parameter biases may partly be a result of the GEFSRv12's t2m bias being attributed to snow cover (Guan et al. 2022), where CAPE and CIN are often negligible. Additionally, only two temperature fields are shown in Fig. 4, which is insufficient to capture the full vertical profile that yields CAPE and CIN calculations.

Further rank histograms were calculated for differing seasons and forecast leads (not shown). Underdispersion dominates in week 1 for all seasons, as expected with a small ensemble. The week 3 and week 4 rankings generally resemble those in week 2. Seasonally, the largest t2m biases are found during the winter and fall, which again supports the results found in Guan et al. (2022) regarding temperatures and snow cover. For t700, the warm bias is consistent year-round. Slight underdispersion is evident for SBCAPE and SRH during the spring, while the other two convective parameters remain well-dispersed. However, given the poor scores in the deterministic metrics for composite convective parameters and dominance of zero values, the degree of dispersion is a less useful discriminator for performance for these fields.

For further comparison between variables, RPSS (Weigel et al. 2007) for tercile categories—being below the 33rd percentile, between the 33rd and 66th percentiles, and above the 66th percentile—is presented in Fig. 5 for differing temporal averages. The fair RPSS (Ferro 2014), which adjusts for smaller ensemble sizes, is used in the calculation. Compared to the results for ACC and RMSE, decreasing the temporal window tends to have a minimal impact on skill. GEFSRv12 convective parameters continue to show an overall poorer performance. For daily means, the inflection point in the mean RPSS where the minimum skill relative to climatology is achieved and generally remains constant is reached by the middle to end of forecast week 2. However, for composite convective parameters (Figs. 5m–p) and SBCIN (Fig. 5l), there is virtually no skill on average relative to climatology at any lead

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 07:02 PM UTC

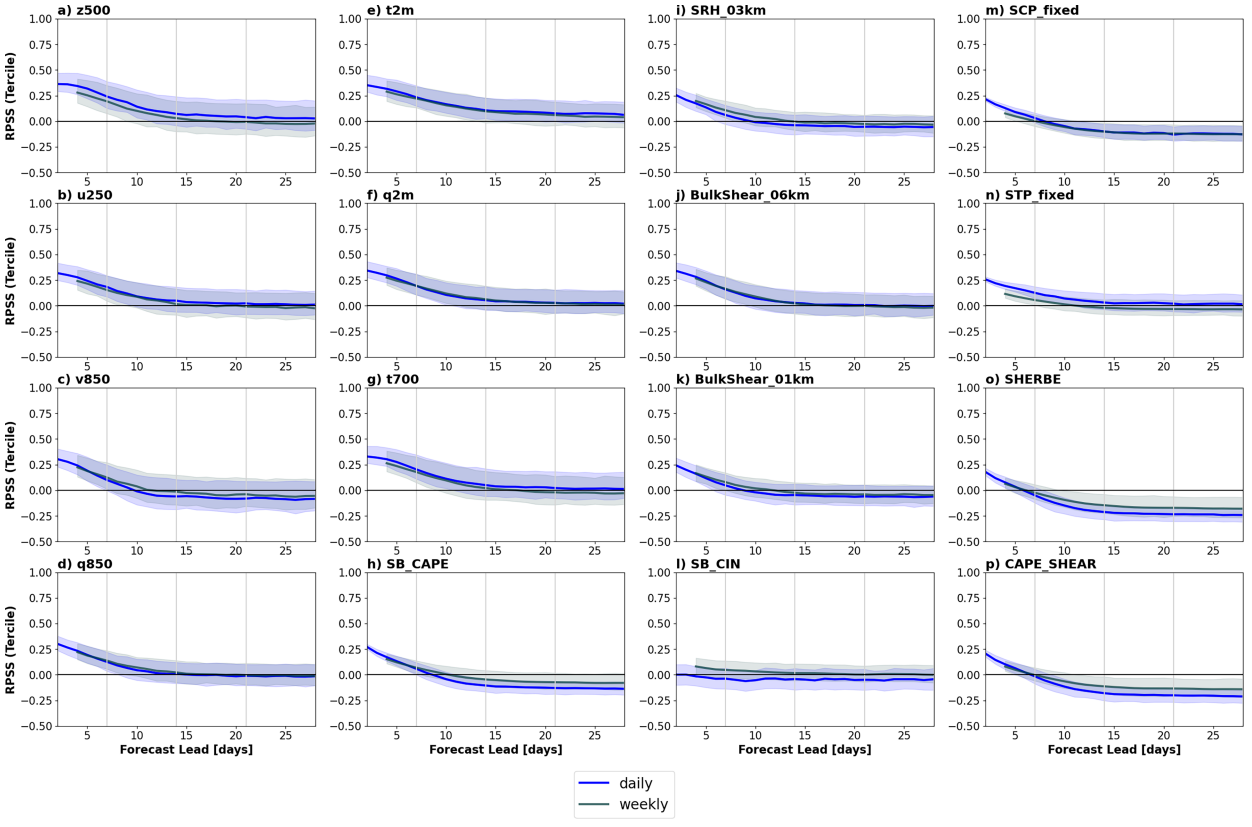

FIG. 5. As in Figs. 1 and 2, but for mean tercile ranked probabilistic skill scores against lead-dependent climatology for forecasts during all seasons.

beyond week 1, with the exception of STP-fixed (Fig. 5n). Even using 75th percentile, RPSS suggests little in the way of forecast opportunities for composite convective parameters. Seasonal RPSS calculations were also performed as with RMSE and ACC, yielding similar results (not shown) with the highest scores in winter and the lowest scores in summer and fall. Additional tests (not shown) using weekly mean persistence and randomly selected forecasts as reference yielded similar results with skill for composite parameters declining by the end of week 1. The slope of RPSS with both persistence and random reference forecasts was steeper than for the model climatology, which shows that the choice of reference climatology can certainly impact the resultant skill scores (Hamill and Juras 2006; Manrique-Suñén et al. 2020).

In summary, for deterministic and probabilistic skill scores, temporal averaging tends to improve skill via deterministic metrics and yields minimal improvement for probabilistic metrics. Individual kinematic and thermodynamic fields outperform vertically integrated and composite convective parameters. Sum aggregation of composite SCS parameters over longer leads may provide clues that are removed when temporally averaging, which are discussed in the following section.

c. *Evaluation of composite parameter weekly sums*

The deterministic and probabilistic measures above suggest that GEFSRv12 can resolve larger-scale kinematic and thermodynamic fields over the CONUS well into week 2, especially when applying temporal generalization. However, Gensini and Tippett (2019) suggested that there is additional promise in using composite convective parameters from the GEFS operational forecasts out to week 2. As mentioned, one issue with using composite parameters is the abundance of zero values that lead to a skewed probability density function at most grid points, especially during less active seasons for SCSs such as winter. To partially address this issue and instead temporally aggregate as opposed to average, we employ weekly summations using a forward window of daily maximum composite convective parameters to investigate the relationship between "favorable" SCS parameter spaces in GEFSRv12 and those in reanalysis.

To define thresholds for statistical tests for the four composite convective parameters, we use weekly sums calculated from the GEFSv12 reanalysis and remove zero values, so there is less skewing toward low values given their frequency. The zero values are only removed for the purposes of establishing higher thresholds for the yes/no verification metrics. No zero values are removed when calculating the verification metrics. Subsequently, we compute the mean and 90th percentile values over all 20 years and all grid points in the domain. Using this method for the mean values, the CAPE–SHEAR threshold is set at 40 000, the SCP-fixed threshold is set at 3, the STP-fixed threshold is set at 0.3, and the

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 07:02 PM UTC

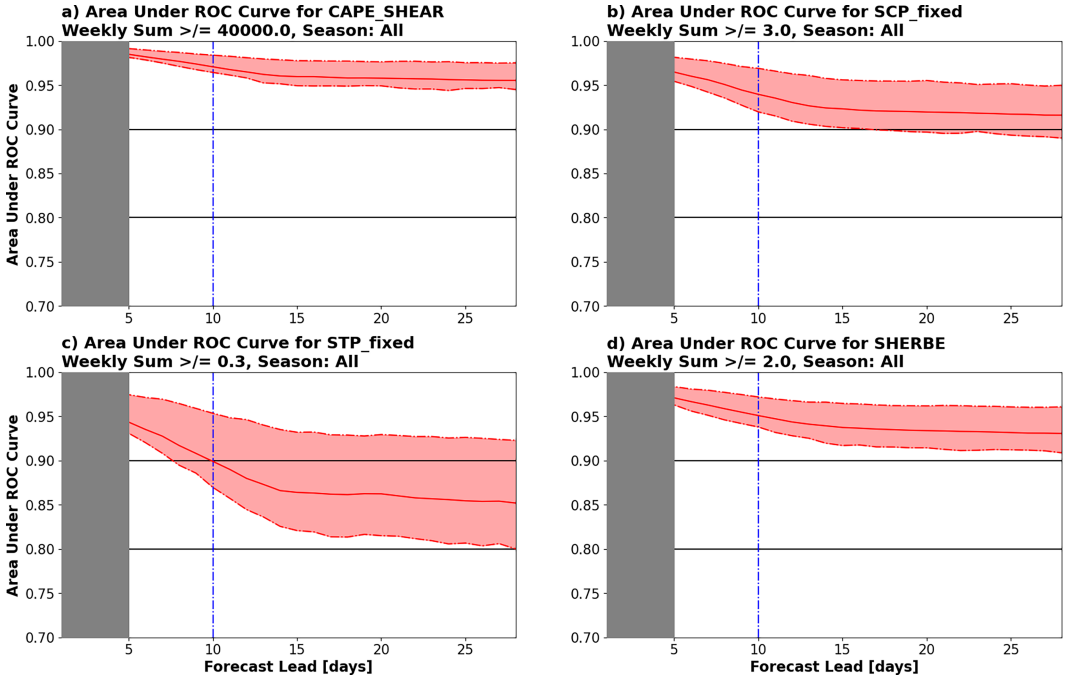

FIG. 6. Climatological AUC for weekly sums of daily maximum: (a) CAPE–SHEAR, (b) SCP-fixed, (c) STP-fixed, and (d) SHERBE for the CONUS domain. The solid red line indicates the mean, while the red dashed–dotted lines indicate the 25th and 75th percentile AUCs at the given leads and the red-shaded region highlights the interquartile range. The blue dashed–dotted line indicates the middle of forecast week 2. Weekly sum exceedance thresholds are indicated in the titles of each subplot. Grayed-out region masks the first 4 days of the forecast, and the plot is shifted to the center of each forward-running weekly period.

SHERBE threshold is set at 2. The STP-fixed threshold is relatively low—especially for a weekly summation—which can be partly attributed to the relative rarity of significant tornadoes. Additionally, the incorporation of convective inhibition in the STP-fixed calculation (Thompson et al. 2003) yields less spatial continuity (smaller regions without zero values) than with the other parameters.

Figure 6 shows the weekly sum climatology of AUC for lead times extending to week 4, which represents the relation between true-positive and false alarm rates. The mean AUC is higher for CAPE–SHEAR, SCP-fixed, and SHERBE at most lead times compared to STP-fixed. As in Gensini and Tippett (2019), the analysis of the ROC curves themselves for differing leads (not shown) suggests that the additional AUC relative to climatology primarily comes from lower false alarm rates at higher probabilities of detection. Seasonal analysis of AUC indicates that more variability exists during the cool seasons of fall and winter, which is expected as CAPE is less available during these seasons.

Another metric for assessing the binary probabilities of exceedance for the weekly sums is the BSS, which is presented in Fig. 7. Here, the annual cycle of forecast skill is emphasized using monthly aggregation instead of seasons, and median scores instead of means are calculated to remove some influence of outliers where the climatological Brier score is very close to zero. Forecasts during the all months consistently perform better than model climatology at all lead times for the four variables, although seasonal variation does remain. Months that have higher incidence of appreciable CAPE (summer) show the greatest skill relative to climatology for CAPE–SHEAR, while the spring months show the best results for the three remaining variables. Focusing on week 2, September is the poorest month relative to climatology for SCP-fixed and STP-fixed, while the winter months are the poorest for CAPE–SHEAR (when CAPE is less available). The time series for SHERBE is the most tightly clustered of the variables, which is perhaps a result of its less reliance on CAPE. Both the seasonal results of AUC and BSS suggest that the availability of instability on a consistent basis is an important factor in determining quality of forecasts using convective parameters. For example, a decrease or increase of 1000 J kg^{-1} CAPE compared to observations may result in a more substantial forecast error during the winter where background climatology is closer to zero versus the summer months (Kirkpatrick et al. 2011).

For the purpose of assessing tails of the weekly sum distribution associated with SCS activity, a 90th percentile threshold BSS evaluation is presented in Fig. 8. Here, weekly sum thresholds increase to 110 000 for CAPE–SHEAR, 7 for SCP-fixed, 0.8 for STP-fixed, and 4 for SHERBE. Skill generally decreases relative to climatology compared to mean threshold values, as one would expect for rarer events. The skill also decays at a faster rate with increasing lead time, particularly for CAPE–SHEAR. Once again, the lowest skill scores relative

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 07:02 PM UTC

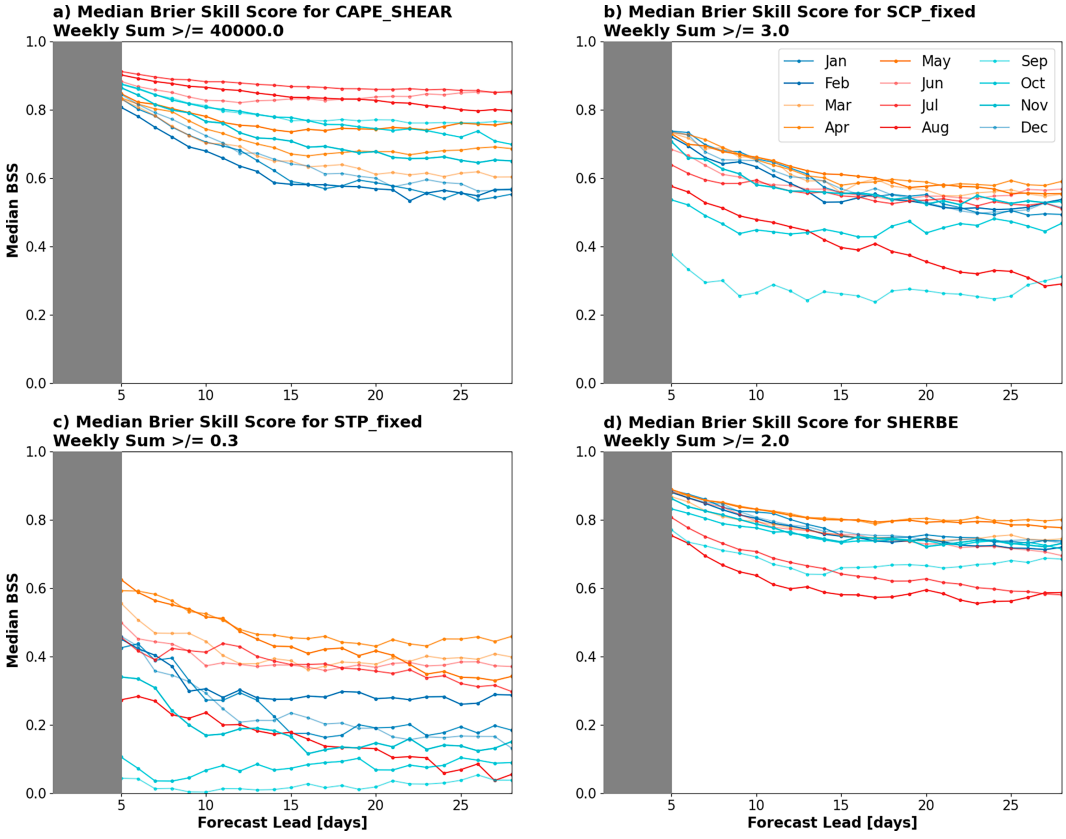

FIG. 7. Monthly climatology of BSSs evaluating weekly sums exceeding the thresholds for the same variables as in Fig. 6. Model climatology is used as the reference forecast. Grayed-out region masks the shifting of the plot to account for the forward-running window as in Fig. 6. Legend for all subplots is located in (b).

to climatology are present during the winter months. Outside of the spring and summer months for extreme values, many of the STP-fixed and SCP-fixed curves decrease below climatological skill via the BSS.

Some caution should be taken in interpreting the results of the AUC and BSS tests above, particularly that the BSS never decreases below the climatological reference forecast, which differs from that of the RPSS shown in Fig. 5. As highlighted in Hamill and Juras (2006), the underlying reference forecast and climatology can be a source of unrealistically high skill in evaluation metrics. Here, we believe that the discrepancy arises from the differing underlying climatologies between the weekly sums and weekly/daily means and, perhaps more broadly, the choice of the model climatology as the reference forecast (Manrique-Suñén et al. 2020). Therefore, it is perhaps more prudent to highlight the slope of each metric with increasing lead, which remains relatively steep through week 2 before flattening in subsequent leads.

While AUC and BSS provide the estimates of forecast skill for the weekly sums, they do not provide the estimates of forecast reliability. For this reason and given the availability of 20 years of data that Gensini and Tippett (2019) did not have access to, we use reliability diagrams to determine whether forecast probabilities for exceeding thresholds are matching the true probabilities from observations.

In Fig. 9, the GEFSRv12 does a reasonable job in matching observed frequencies for exceeding the mean thresholds. Particularly for CAPE–SHEAR, reliability curves for all forecast leads follow the 1:1 (perfect reliability) line to some degree. There is a slight tendency toward underconfidence—higher true frequencies than predicted probabilities—at lower (less than 10%) forecast probabilities for CAPE–SHEAR. There is also notable underconfidence during week 1 for SHERBE, with lesser underconfidence during week 2 and beyond. For SCP-fixed and STP-fixed, there is a tendency for overconfidence (higher predicted probabilities than observed frequencies) at higher probability thresholds greater than 50%. These results are generally reflected when evaluating 90th percentile threshold values in Fig. 10. During week 1, SCP-fixed and STP-fixed show minimal skill via their reliability curves for extreme values, but improve marginally during week 2 and beyond. There remains underprediction during week 1 for SHERBE but close following the perfect reliability line during subsequent weeks. The reliability curves for 90th percentile values during all weeks for CAPE–SHEAR closely follow the 1:1 line.

Notably, all reliability curves for week 2—where deterministic scores decay below climatology—show promising results and despite some deviation from the perfect reliability line,

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 07:02 PM UTC

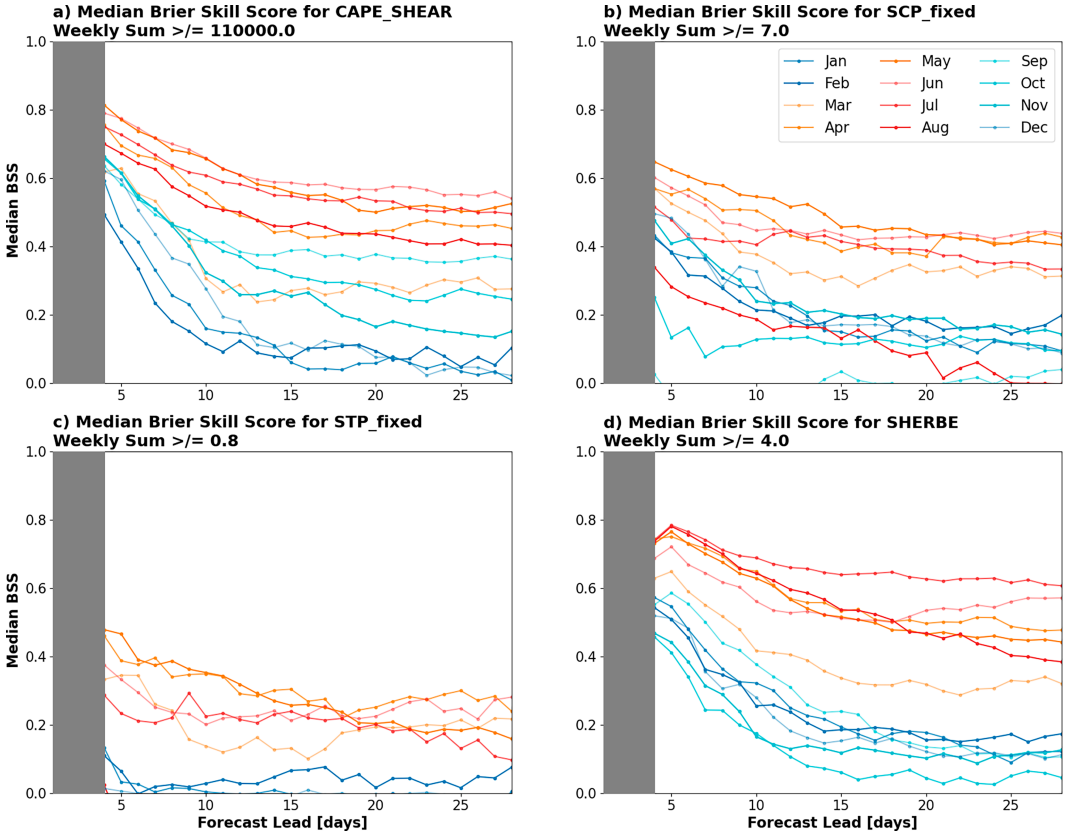

FIG. 8. As in Fig. 7, but for 90th percentile exceedance thresholds instead of mean values.

all variables show improvement from the no skill threshold. This result is consistent for both sets of thresholds tested and for all seasons. The reliability diagrams for the weekly sums show indications that GEFSRv12 is representing probabilities for SCS environments reasonably well when aggregating.

Statistical summation can add additional information that may be removed when averaging (Li and Stechmann 2018)—especially over longer periods such as 1 week—with the reduction in the number of zero values in the raw (nonanomaly) data. It is thus expected that the underlying climatological standard deviation becomes larger in magnitude when using weekly sums of daily maxima versus weekly averages. To further emphasize the usefulness of weekly sums when evaluating convective parameters, climatological anomaly correlation coefficients are again presented in Fig. 11 and compared to the daily max and rolling weekly average from Fig. 1. A notable improvement in ACC skill (values summarized in Table 1) for weekly sums is present for all four composite convective parameters and especially for CAPE–SHEAR. For SCP-fixed, STP-fixed, and CAPE–SHEAR, while daily maxes and even weekly averages struggle to achieve synoptic or climatological skill into week 2, weekly summations extend skill for mean ACC into the early or even middle of week 2. For SHERBE, while weekly averaging showed some promise, weekly summation further extends appreciable skill in the mean to day 10. The 75th percentile ACC for all four parameters extends toward the end of week 2 for climatological skill, thus representing additional increases in skill using weekly summations compared to weekly averages.

As a final illustration of the improvements in skill offered by temporal averaging or aggregation, the mean weekly ACC for weeks 1 (days 1–7), 2 (days 8–14), 3 (days 15–21), and 4 (days 22–28) is calculated for all variables used in the study and summarized in Table 1. For kinematic, circulation, and nonvertically integrated thermodynamic variables, the average ACC for week 1 lies between 0.7 and 0.9 for weekly means and decreases to 0.4–0.6 for week 2. The 75th percentile ACC (not shown) eclipses synoptic skill (0.6) thresholds during week 2 for several of these variables. BS06 also yields appreciable skill into week 2 when using a weekly average. For the remainder of variables tested, while weekly averaging does improve skill over daily means, it does not generally exceed the synoptic or climatological (0.5) threshold. However, for weekly summations, the average week 2 ACC for three of the composite convective parameters rises close to the climatological skill threshold. The results for weeks 3 and 4 are generally unfavorable for all variables, indicating the lack of predictability (Miller and Gensini 2023) at this range.

Weekly summations of convective parameters yield promising indications via a number of skill metrics compared to

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 07:02 PM UTC

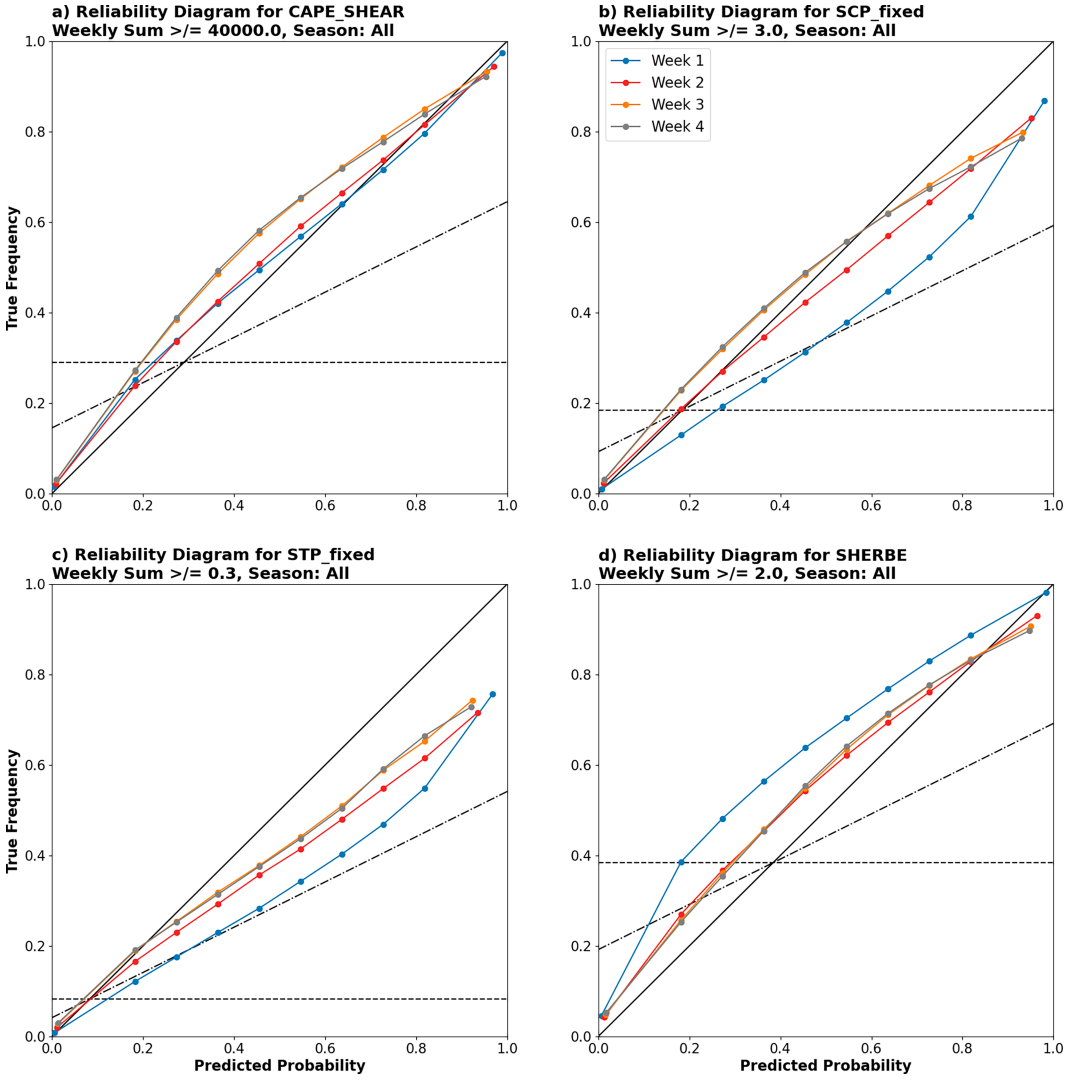

FIG. 9. CONUS reliability diagrams for the same variables as in Figs. 6 and 7 for exceedance probabilities at varying forecast leads during all seasons. The black solid line indicates the 1:1 line of perfect reliability. The black dashed line indicates climatological probabilities for exceeding the threshold at any given grid point, and the black dashed–dotted line indicates the no skill delineation between perfect reliability and climatology. Reliability curves are calculated using 10 probability bins. Legend for all subplots is located in (b).

weekly averages. The results of this section further suggest that work into predicting SCSs at extended ranges using model variables should focus on the early portion of the subseasonal time scale—as meaningful skill rarely exists beyond that lead time—at least in the GEFSRv12. Exceptions may exist (Gensini et al. 2019; Miller and Gensini 2023), but when considering an average across a large number of forecasts and potential operational implications, the most potential for expansion of current SCS forecasts exists in week 2.

4. Conclusions and discussion

Prediction of severe convective storms (SCSs) beyond week 1 has been of interest to the meteorological community in recent years. Studies dedicated to SCS forecasting outside of forecast week 1 have been primarily focused on statistical work (Barrett and Gensini 2013; Gensini and Marinaro 2016; Allen et al. 2018; Gensini et al. 2019), and studies that have attempted to forecast SCSs at these lead times have not evaluated the variables used as proxies (Gensini and Tippett 2019; Lee et al. 2021; Miller et al. 2022). Thus, this work attempts to establish a baseline climatology for forecast parameters. Prior literature studies examining forecasts for SCS environments have been limited by smaller sample sizes and short temporal windows of available model data (Gensini and Tippett 2019). Therefore, recently released GEFSv12 reforecast (GEFSRv12) data (Guan et al. 2022; Zhou et al. 2022) over the 20-yr period from 2000 through 2019 are used over the CONUS domain to

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 07:02 PM UTC

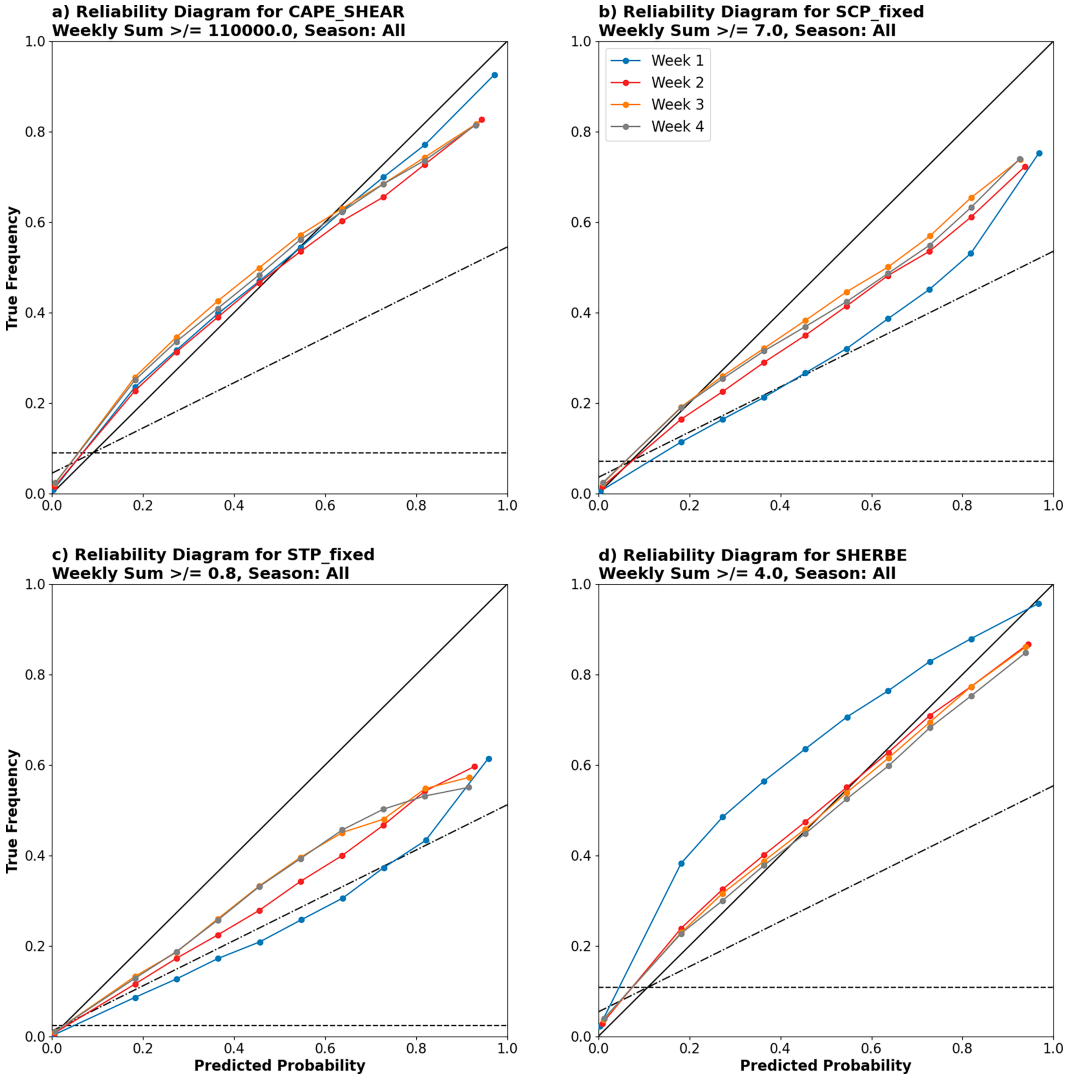

FIG. 10. As in Fig. 9, but for 90th percentile exceedance thresholds instead of mean values.

provide a multidecadal dataset of forecasts. A variety of deterministic and probabilistic metrics are used to evaluate 35-day reforecasts issued weekly, in addition to a novel use of weekly summations of composite SCS parameters to attempt to extend their use into forecast week 2.

Results suggest that variables concerning synoptic-scale variability across the CONUS can be reasonably resolved into week 2 by the GEFSRv12, although ensemble dispersion suffers during week 1 given the small size of the ensemble. Forecasts of opportunity (Gensini et al. 2019; Mariotti et al. 2020; Miller and Gensini 2023) exist into week 3 via anomaly correlation analysis and ranked probability skill scores for some variables such as 500-hPa geopotential height (z500) and 2-m temperature (t2m). Biases exist for thermodynamic variables that have been documented previously (Guan et al. 2022), although 700-hPa temperature (t700) shows the opposite bias (warm) to t2m. Concurrent work (Miller and Gensini 2023) determined favorable global weather regimes for these higher skill windows, particularly for z500. The findings surrounding z500 additionally support the exploratory work of Miller et al. (2020) using WRs to skillfully predict SCS activity at longer time scales.

The seasonal cycle of forecast skill is evident, with kinematic and thermodynamic fields showing longer windows of skill in the winter and early spring versus summer. Additionally, temporal averaging of fields (van Straaten et al. 2020) into 7-day rolling means adds some additional predictability as evidenced by deterministic and probabilistic verification measures. However, care should be taken when smoothing variables that vary on shorter time scales such as composite SCS parameters, as this tends to remove variability associated with these fields. As this work is meant to act as a precursor for later calibrated forecast and machine learning experiments, this work suggests that only single-level kinematic and thermodynamic fields from GEFSRv12 should be utilized as inputs for forecasts at leads greater than 1 week and that

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 07:02 PM UTC

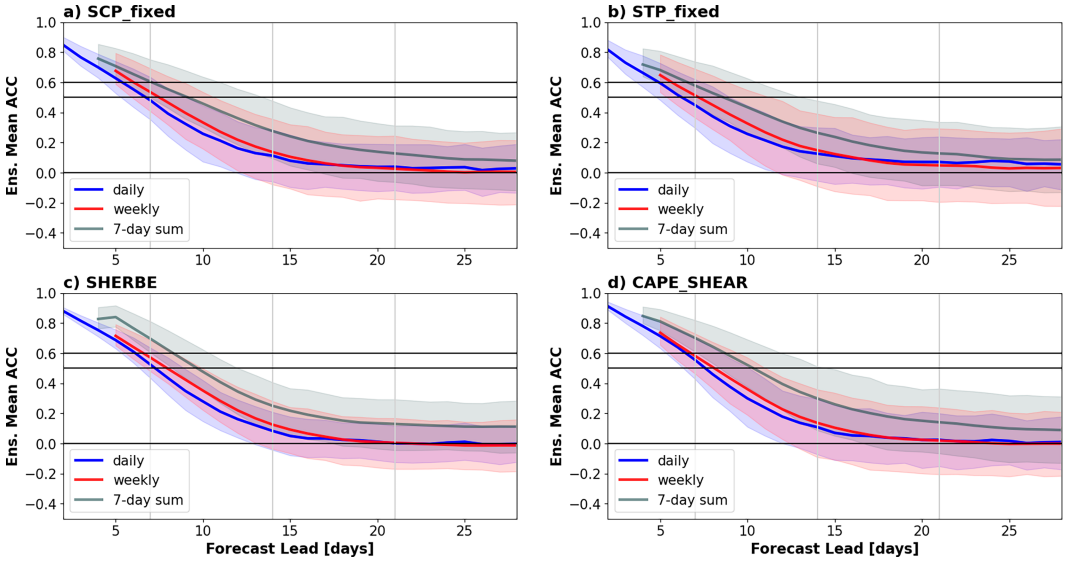

FIG. 11. As in Fig. 1, but only for composite convective parameters and including ACC for 7-day summations. The red line/shading now represents the 7-day rolling mean, and the gray line/shading represents the 7-day summation.

some degree of temporal averaging is likely to assist in forecasts beyond week 1.

For SCS composite parameters such as SCP-fixed and STP-fixed, skill via traditional deterministic and probabilistic metrics rarely exists beyond week 1. Vertically integrated fields such as CAPE and CIN also struggle. Errors in boundary layer parameterization (Cohen et al. 2017) may be a leading cause of these difficulties, along with the vertical resolution of the ensemble guidance. For SCS composite parameters, the mere fact that multiple variables are being combined together could be responsible for faster error growth, in addition to their design in prediction of mesoscale features that have shorter intrinsic predictability. For these reasons, a method using aggregated sums of daily maximum composite parameter values over longer temporal windows is developed in this work and tested with promising results from some probabilistic skill metrics, such as reliability diagrams and Brier skill scores (BSSs), which showed improvement from climatological forecasts. Weekly summation also provides additional skill compared to weekly averaging for deterministic measures (Fig. 11).

A number of caveats exist in this analysis. To start, all fields in this work are compared using model reanalysis as the observations, which itself may contain biases and errors compared to actual observations (Taszarek et al. 2020; Li et al. 2020). However, the GEFSv12 reanalysis is internally consistent with

TABLE 1. Average ACC during forecast weeks 1–4 (W1–4) for the 16 chosen variables. Columns represent weekly averages and weekly summations. The bolded text represents the ACC exceeding the climatological threshold at W2 or later.

Variable	W1 ACC mean	W1 sum ACC mean	W2 ACC mean	W2 sum ACC mean	W3 ACC mean	W3 sum ACC mean	W4 ACC mean	W4 sum ACC mean
z500	0.866	x	**0.522**	x	0.215	x	0.095	x
u250	0.832	x	**0.501**	x	0.220	x	0.115	x
v850	0.779	x	0.401	x	0.130	x	0.044	x
q850	0.732	x	0.416	x	0.185	x	0.112	x
t2m	0.838	x	**0.540**	x	0.270	x	0.160	x
q2m	0.811	x	**0.504**	x	0.247	x	0.158	x
t700	0.857	x	**0.524**	x	0.217	x	0.107	x
SBCAPE	0.680	x	0.384	x	0.176	x	0.099	x
SRH	0.705	x	0.377	x	0.142	x	0.069	x
BS06	0.811	x	0.480	x	0.204	x	0.107	x
BS01	0.682	x	0.343	x	0.122	x	0.061	x
SBCIN	0.640	x	0.349	x	0.158	x	0.107	x
SCP-fixed	0.545	0.675	0.209	0.412	0.019	0.175	−0.019	0.097
STP-fixed	0.523	0.650	0.211	0.391	0.039	0.171	0.008	0.1
SHERBE	0.582	0.825	0.213	0.421	−0.001	0.163	−0.036	0.117
CAPE–SHEAR	0.596	0.706	0.223	0.464	0.013	0.190	−0.024	0.108

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 07:02 PM UTC

the reforecast data. The list of variables examined in this study is by no means comprehensive, nor is the small size of the ensemble reforecast, which should be compared with operational NWP ensembles. The thresholds identified in the weekly sum analysis for convective variables are based on domainwide averages. Regional dependence for many of these variables should be further explored and rolling climatologies may also be used to better represent the seasonal cycle versus rigidly defined seasons.

More broadly, the use of model variables as proxies for SCS activity (Gensini and Tippett 2019; Lee et al. 2021) is not the primary objective of the study. Rather, this work represents an intermediate step between raw model output and using proxy parameters to estimate SCS activity that occurs. It is an attempt to narrow down the methods and variables used for SCS forecasting that are useful beyond week 1 and to reduce the "garbage in, garbage out" problem (Hall 2014) that often affects forecasts that use postprocessed data from models.

Despite these issues and considerations, this evaluation provides a more comprehensive climatology of forecast parameters relevant for SCS prediction than previously explored in the literature. The results presented act as a baseline for subsequent work in improving SCS forecasts time scales beyond week 1 using GEFS, especially during week 2 when meaningful skill does often exist. Ongoing and proposed work using machine learning at longer leads (Hill et al. 2023) to both classify and predict weather regimes favorable for SCSs (Miller et al. 2020) may use such evaluation studies to better understand variables that are useful at these time scales. The reforecast evaluation could also expand to other regions, particularly the North Pacific, where the variability of the jet stream has shown links to the North American pattern (Gensini et al. 2019; Winters 2021). Future forecast experiments at subseasonal leads for SCSs may also incorporate information about model-predicted or observed MJO and other subseasonal variability helpful in the identification of higher skill periods (Miller and Gensini 2023), which is beyond the scope of this paper. We hope that this work can complement such studies in efforts to expand prediction of SCSs beyond current capabilities.

Acknowledgments. The authors thank Thomas Galarneau, Thomas Hamill, and two anonymous reviewers for providing their helpful comments that improved the manuscript from its original version. Funding was provided by the NOAA/Office of Oceanic and Atmospheric Research under NOAA—University of Oklahoma Cooperative Agreement NA21OAR4320204, U.S. Department of Commerce.

Data availability statement. Datasets used in this manuscript are publicly available. GEFSv12 reforecast data are available via the Registry of Open Data at Amazon Web Services (AWS) at https://registry.opendata.aws/noaa-gefs-reforecast/. GEFSv12 reanalysis data used in this study are available from NOAA's Environmental Modeling Center via FTP download at ftp://ftp.emc.ncep.noaa.gov/GEFSv12/reanalysis/FV3_reanalysis.

APPENDIX

Supplemental Equations

a. Equations of convective composite parameters

The equations for the supercell composite parameter (SCP) and significant tornado parameter (STP) are given in Thompson et al. (2003, 2012) and are modified appropriately for their fixed-layer versions using 0–3- or 0–1-km SRH (SRH03 or SRH01), SBCAPE, surface-based lifted condensation levels (SBLCLs), and 0–6-km bulk shear (BS06) instead of their effective or mixed-layer counterparts,

$$\mathrm{SCP}_{\mathrm{fixed}} = \frac{\mathrm{MUCAPE}}{1000\,\mathrm{J\,kg^{-1}}} \times \frac{\mathrm{BS06}}{20\,\mathrm{m\,s^{-1}}} \times \frac{\mathrm{SRH03}}{50\,\mathrm{m^2\,s^{-2}}}, \tag{A1}$$

$$\mathrm{STP}_{\mathrm{fixed}} = \frac{\mathrm{SBCAPE}}{1500\,\mathrm{J\,kg^{-1}}} \times \frac{(2000 - \mathrm{SBLCL})}{1000\,\mathrm{m}} \times \frac{\mathrm{BS06}}{20\,\mathrm{m\,s^{-1}}} \times \frac{\mathrm{SRH01}}{150\,\mathrm{m^2\,s^{-2}}} \times \frac{(200 + \mathrm{SBCIN})}{150\,\mathrm{J\,kg^{-1}}}. \tag{A2}$$

For the SCP formulation, the BS06 term is set to zero when BS06 is less than 10 m s^{-1} and becomes 1 when values are greater than 20 m s^{-1}. For the STP formulation, the SBLCL term is set to 1 when the SBLCL is less than 1000 m AGL and set to 0 when the SBLCL is greater than 2000 m AGL. The SBCIN term is set to 1 when SBCIN is greater than -50 J kg^{-1} and set to 0 when SBCIN is less than -200 J kg^{-1}. Finally, the BS06 term is set at 1.5 when BS06 is greater than 30 m s^{-1} and set to 0 when BS06 is less than 12.5 m s^{-1}.

Meanwhile, the equation for SHERBE is given in Sherburn and Parker (2014) and is as follows:

$$\mathrm{SHERBE} = \frac{\mathrm{BSE}}{27\,\mathrm{m\,s^{-1}}} \times \frac{\mathrm{LLLR}}{5.2\,\mathrm{K\,km^{-1}}} \times \frac{\mathrm{LR75}}{5.6\,\mathrm{K\,km^{-1}}}. \tag{A3}$$

Here, BSE represents the effective bulk wind shear (m s^{-1}), LLLR represents the low-level lapse rate (K km^{-1}), and LR75 represents the 700–500-hPa lapse rate (K km^{-1}).

b. Equations of verification metrics

The equations for anomaly correlation coefficient (ACC) and root-mean-square error (RMSE) of the ensemble mean at time t are as follows and are provided in a modified version from that in Joliffe and Stephenson (2012),

$$\mathrm{ACC} = \frac{\sum_{m=1}^{M} (\bar{f}_m - c_m)(o_m - c_m)}{\sqrt{\sum_{m=1}^{M} (\bar{f}_m - c_m)^2} \sqrt{\sum_{m=1}^{M} (o_m - c_m)^2}}, \tag{A4}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{M} \sum_{m=1}^{M} (\bar{f}_m - o_m)^2}. \tag{A5}$$

Here, m represents the mth grid point, M represents the total number of grid points (71 × 171), $\bar{f}$ represents the

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 07:02 PM UTC

forecast value (GEFSv12 reforecast ensemble mean) at time t, c represents the climatological mean value at time t, and o represents the observed (GEFSv12 reanalysis) value at time t.

For the ensemble consistency, the mean-square error at time t is simply the RMSE without the square root but is modified to account for the small ensemble size as in Eckel and Mass (2005). Ensemble variance is also calculated using this adjustment and using the individual ensemble members:

$$\mathrm{MSE} = \left(\frac{n}{n+1}\right)\frac{1}{M}\sum_{m=1}^{M}(\bar{f}_m - o_m)^2, \tag{A6}$$

$$\mathrm{Var}_e = \frac{1}{M}\sum_{m=1}^{M}\left[\frac{1}{n-1}\sum_{i=1}^{n}(f_{m,i} - \bar{f}_m)\right]. \tag{A7}$$

Here, i represents the ith ensemble member and $f_{m,i}$ is the forecast value of the ith ensemble member at the mth grid point at the given time t.

For probabilistic skill scores of a given forecast–observation pair at a time t, equations for the ranked probability score are given in Weigel et al. (2007) for both forecasts and climatology. Scores are calculated relative to their local climatologies before being pooled together in summary statistics,

$$\mathrm{RPS} = \sum_{k=1}^{K}(\mathbf{F}_k - \mathbf{O}_k)^2, \tag{A8}$$

$$\mathrm{RPS}_{\mathrm{clim}} = \sum_{k=1}^{K}(\mathbf{CF}_k - \mathbf{O}_k)^2. \tag{A9}$$

Here, k represents the kth category, K represents the total number of categories, $\mathbf{F}$ is the forecast vector, $\mathbf{CF}_k$ is the climatological forecast vector, and $\mathbf{O}$ is the observation vector. The Brier score simply corresponds to the binary case where $K = 2$. The $\mathbf{CF}_k$ in our study refers to the lead-dependent climatology as described in section 2b. The corresponding ranked probability skill score (RPSS) is calculated as follows (Weigel et al. 2007; Wilks 2011):

$$\mathrm{RPSS} = 1 - \frac{\langle \mathrm{RPS} \rangle}{\langle \mathrm{RPS}_{\mathrm{clim}} \rangle}. \tag{A10}$$

Here, $\langle \mathrm{RPS} \rangle$ is the average RPS across all forecast–observation vectors, while $\langle \mathrm{RPS}_{\mathrm{clim}} \rangle$ is the average RPS across all climatological forecast–observation vectors. For more detailed information on these scores, see Epstein (1969), Weigel et al. (2007), Wilks (2011), and Joliffe and Stephenson (2012). The correction used for adjusting these scores for the ensemble size is detailed in Ferro (2014).

REFERENCES

Allen, J. T., M. J. Molina, and V. A. Gensini, 2018: Modulation of annual cycle of tornadoes by El Niño–Southern Oscillation. *Geophys. Res. Lett.*, **45**, 5708–5717, https://doi.org/10.1029/2018GL077482.

Baggett, C. F., K. M. Nardi, S. J. Childs, S. N. Zito, E. A. Barnes, and E. D. Maloney, 2018: Skillful subseasonal forecasts of weekly tornado and hail activity using the Madden-Julian Oscillation. *J. Geophys. Res. Atmos.*, **123**, 12 661–12 675, https://doi.org/10.1029/2018JD029059.

Barrett, B. S., and V. A. Gensini, 2013: Variability of central United States April–May tornado day likelihood by phase of the Madden-Julian Oscillation. *Geophys. Res. Lett.*, **40**, 2790–2795, https://doi.org/10.1002/grl.50522.

——, and B. N. Henley, 2015: Intraseasonal variability of hail in the contiguous United States: Relationship to the Madden–Julian oscillation. *Mon. Wea. Rev.*, **143**, 1086–1103, https://doi.org/10.1175/MWR-D-14-00257.1.

Baxter, M. A., G. M. Lackmann, K. M. Mahoney, T. E. Workoff, and T. M. Hamill, 2014: Verification of quantitative precipitation reforecasts over the southeastern United States. *Wea. Forecasting*, **29**, 1199–1207, https://doi.org/10.1175/WAF-D-14-00055.1.

Bröcker, J., and L. A. Smith, 2007: Increasing the reliability of reliability diagrams. *Wea. Forecasting*, **22**, 651–661, https://doi.org/10.1175/WAF993.1.

Büeler, D., L. Ferranti, L. Magnusson, J. F. Quinting, and C. M. Grams, 2021: Year-round sub-seasonal forecast skill for Atlantic–European weather regimes. *Quart. J. Roy. Meteor. Soc.*, **147**, 4283–4309, https://doi.org/10.1002/qj.4178.

Buizza, R., and M. Leutbecher, 2015: The forecast skill horizon. *Quart. J. Roy. Meteor. Soc.*, **141**, 3366–3382, https://doi.org/10.1002/qj.2619.

Carbin, G. W., M. K. Tippett, S. P. Lillo, and H. E. Brooks, 2016: Visualizing long-range severe thunderstorm environment guidance from CFSv2. *Bull. Amer. Meteor. Soc.*, **97**, 1021–1031, https://doi.org/10.1175/BAMS-D-14-00136.1.

Clark, A. J., W. A. Gallus Jr, M. Xue, and F. Kong, 2010: Growth of spread in convection-allowing and convection-parameterizing ensembles. *Wea. Forecasting*, **25**, 594–612, https://doi.org/10.1175/2009WAF2222318.1.

Cohen, A. E., S. M. Cavallo, M. C. Coniglio, H. E. Brooks, and I. L. Jirak, 2017: Evaluation of multiple planetary boundary layer parameterization schemes in southeast U.S. cold season severe thunderstorm environments. *Wea. Forecasting*, **32**, 1857–1884, https://doi.org/10.1175/WAF-D-16-0193.1.

Craig, G. C., and Coauthors, 2021: Waves to weather: Exploring the limits of predictability of weather. *Bull. Amer. Meteor. Soc.*, **102**, E2151–E2164, https://doi.org/10.1175/BAMS-D-20-0035.1.

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

Eckel, F. A., and C. F. Mass, 2005: Aspects of effective mesoscale, short-range ensemble forecasting. *Wea. Forecasting*, **20**, 328–350, https://doi.org/10.1175/WAF843.1.

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.*, **8**, 985–987, https://doi.org/10.1175/1520-0450(1969)008<0985:ASSFPF>2.0.CO;2.

Ferro, C. A., 2014: Fair scores for ensemble forecasts. *Quart. J. Roy. Meteor. Soc.*, **140**, 1917–1923, https://doi.org/10.1002/qj.2270.

Gensini, V. A., and A. Marinaro, 2016: Tornado frequency in the United States related to global relative angular momentum. *Mon. Wea. Rev.*, **144**, 801–810, https://doi.org/10.1175/MWR-D-15-0289.1.

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 07:02 PM UTC

——, and J. T. Allen, 2018: U.S. hail frequency and the Global Wind Oscillation. *Geophys. Res. Lett.*, **45**, 1611–1620, https://doi.org/10.1002/2017GL076822.

——, and L. B. De Guenni, 2019: Environmental covariate representation of seasonal U.S. tornado frequency. *J. Appl. Meteor. Climatol.*, **58**, 1353–1367, https://doi.org/10.1175/JAMC-D-18-0305.1.

——, and M. K. Tippett, 2019: Global Ensemble Forecast System (GEFS) predictions of days 1–15 U.S. tornado and hail frequencies. *Geophys. Res. Lett.*, **46**, 2922–2930, https://doi.org/10.1029/2018GL081724.

——, D. Gold, J. T. Allen, and B. S. Barrett, 2019: Extended U.S. tornado outbreak during late May 2019: A forecast of opportunity. *Geophys. Res. Lett.*, **46**, 10 150–10 158, https://doi.org/10.1029/2019GL084470.

——, B. S. Barrett, J. T. Allen, D. Gold, and P. Sirvatka, 2020: The Extended-Range Tornado Activity Forecast (ERTAF) project. *Bull. Amer. Meteor. Soc.*, **101**, E700–E709, https://doi.org/10.1175/BAMS-D-19-0188.1.

Goutham, N., R. Plougonven, H. Omrani, S. Parey, P. Tankov, A. Tantet, P. Hitchcock, and P. Drobinski, 2022: How skillful are the European subseasonal predictions of wind speed and surface temperature? *Mon. Wea. Rev.*, **150**, 1621–1637, https://doi.org/10.1175/MWR-D-21-0207.1.

Grams, J. S., R. L. Thompson, D. V. Snively, J. A. Prentice, G. M. Hodges, and L. J. Reames, 2012: A climatology and comparison of parameters for significant tornado events in the United States. *Wea. Forecasting*, **27**, 106–123, https://doi.org/10.1175/WAF-D-11-00008.1.

Guan, H., and Coauthors, 2022: GEFSv12 reforecast dataset for supporting subseasonal and hydrometeorological applications. *Mon. Wea. Rev.*, **150**, 647–665, https://doi.org/10.1175/MWR-D-21-0245.1.

Hall, A., 2014: Projecting regional change. *Science*, **346**, 1461–1462, https://doi.org/10.1126/science.aaa0629.

Hamill, T. M., 1997: Reliability diagrams for multicategory probabilistic forecasts. *Wea. Forecasting*, **12**, 736–741, https://doi.org/10.1175/1520-0434(1997)012<0736:RDFMPF>2.0.CO;2.

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

——, and J. Juras, 2006: Measuring forecast skill: Is it real skill or is it the varying climatology? *Quart. J. Roy. Meteor. Soc.*, **132**, 2905–2923, https://doi.org/10.1256/qj.06.25.

——, J. S. Whitaker, and S. L. Mullen, 2006: Reforecasts: An important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc.*, **87**, 33–46, https://doi.org/10.1175/BAMS-87-1-33.

——, R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Mon. Wea. Rev.*, **136**, 2620–2632, https://doi.org/10.1175/2007MWR2411.1.

——, G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau, Y. Zhu, and W. Lapenta, 2013: NOAA's second-generation global medium-range ensemble reforecast dataset. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565, https://doi.org/10.1175/BAMS-D-12-00014.1.

——, and Coauthors, 2022: The reanalysis for the Global Ensemble Forecast System, version 12. *Mon. Wea. Rev.*, **150**, 59–79, https://doi.org/10.1175/MWR-D-21-0023.1.

Herman, G. R., and R. S. Schumacher, 2016: Extreme precipitation in models: An evaluation. *Wea. Forecasting*, **31**, 1853–1879, https://doi.org/10.1175/WAF-D-16-0093.1.

——, E. R. Nielsen, and R. S. Schumacher, 2018: Probabilistic verification of Storm Prediction Center convective outlooks. *Wea. Forecasting*, **33**, 161–184, https://doi.org/10.1175/WAF-D-17-0104.1.

Hill, A. J., G. R. Herman, and R. S. Schumacher, 2020: Forecasting severe weather with random forests. *Mon. Wea. Rev.*, **148**, 2135–2161, https://doi.org/10.1175/MWR-D-19-0344.1.

——, R. S. Schumacher, and I. L. Jirak, 2023: A new paradigm for medium-range severe weather forecasts: Probabilistic random forest–based predictions. *Wea. Forecasting*, **38**, 251–272, https://doi.org/10.1175/WAF-D-22-0143.1.

Hitchens, N. M., and H. E. Brooks, 2014: Evaluation of the Storm Prediction Center's convective outlooks from day 3 through day 1. *Wea. Forecasting*, **29**, 1134–1142, https://doi.org/10.1175/WAF-D-13-00132.1.

Hu, X.-M., J. Park, T. Supinie, N. A. Snook, M. Xue, K. A. Brewster, J. Brotzge, and J. R. Carley, 2022: Diagnosing near-surface model errors with candidate physics parameterization schemes for the multiphysics Rapid Refresh Forecast System (RRFS) ensemble during winter over the northeastern United States and southern Great Plains. *Mon. Wea. Rev.*, **151**, 39–61, https://doi.org/10.1175/MWR-D-22-0085.1.

Joliffe, I. T., and D. B. Stephenson, 2012: *Forecast Verification: A Practitioner's Guide in Atmospheric Science.* 2nd ed. Wiley, 304 pp.

Kiladis, G. N., J. Dias, K. H. Straub, M. C. Wheeler, S. N. Tulich, K. Kikuchi, K. M. Weickmann, and M. J. Ventrice, 2014: A comparison of OLR and circulation-based indices for tracking the MJO. *Mon. Wea. Rev.*, **142**, 1697–1715, https://doi.org/10.1175/MWR-D-13-00301.1.

Kirkpatrick, C., E. W. Mccaul, and C. Cohen, 2011: Sensitivities of simulated convective storms to environmental CAPE. *Mon. Wea. Rev.*, **139**, 3514–3532, https://doi.org/10.1175/2011MWR3631.1.

Lee, S. K., H. Lopez, D. Kim, A. T. Wittenberg, and A. Kumar, 2021: A Seasonal Probabilistic Outlook for Tornadoes (SPOTter) in the contiguous United States based on the leading patterns of large-scale atmospheric anomalies. *Mon. Wea. Rev.*, **149**, 901–919, https://doi.org/10.1175/MWR-D-20-0223.1.

Lepore, C., M. K. Tippett, and J. T. Allen, 2018: CFSv2 monthly forecasts of tornado and hail activity. *Wea. Forecasting*, **33**, 1283–1297, https://doi.org/10.1175/WAF-D-18-0054.1.

Li, F., D. R. Chavas, K. A. Reed, and D. T. Dawson, 2020: Climatology of severe local storm environments and synoptic-scale features over North America in ERA5 reanalysis and CAM6 simulation. *J. Climate*, **33**, 8339–8365, https://doi.org/10.1175/JCLI-D-19-0986.1.

Li, Y., and S. N. Stechmann, 2018: Spatial and temporal averaging windows and their impact on forecasting: Exactly solvable examples. *Math. Climate Wea. Forecasting*, **4**, 23–49, https://doi.org/10.1515/mcwf-2018-0002.

Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, **21A**, 289–307, https://doi.org/10.1111/j.2153-3490.1969.tb00444.x.

Madden, R. A., and P. R. Julian, 1972: Description of global-scale circulation cells in the tropics with a 40–50 day period. *J. Atmos. Sci.*, **29**, 1109–1123, https://doi.org/10.1175/1520-0469(1972)029<1109:DOGSCC>2.0.CO;2.

Manrique-Suñén, A., N. Gonzalez-Reviriego, V. Torralba, N. Cortesi, and F. J. Doblas-Reyes, 2020: Choices in the verification of S2S forecasts and their implications for climate services. *Mon. Wea. Rev.*, **148**, 3995–4008, https://doi.org/10.1175/MWR-D-20-0067.1.

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 07:02 PM UTC

Mariotti, A., and Coauthors, 2020: Windows of opportunity for skillful forecasts subseasonal to seasonal and beyond. *Bull. Amer. Meteor. Soc.*, **101**, E608–E625, https://doi.org/10.1175/BAMS-D-18-0326.1.

Markowski, P. M., 2020: What is the intrinsic predictability of tornadic supercell thunderstorms? *Mon. Wea. Rev.*, **148**, 3157–3180, https://doi.org/10.1175/MWR-D-20-0076.1.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.*, **30**, 291–303.

Melhauser, C., and F. Zhang, 2012: Practical and intrinsic predictability of severe and convective weather at the mesoscales. *J. Atmos. Sci.*, **69**, 3350–3371, https://doi.org/10.1175/JAS-D-11-0315.1.

Miller, D. E., and V. A. Gensini, 2023: GEFSv12 high- and low-skill day-10 tornado forecasts. *Wea. Forecasting*, **38**, 1195–1207, https://doi.org/10.1175/WAF-D-22-0122.1.

——, Z. Wang, R. J. Trapp, and D. S. Harnos, 2020: Hybrid prediction of weekly tornado activity out to week 3: Utilizing weather regimes. *Geophys. Res. Lett.*, **47**, e2020GL087253, https://doi.org/10.1029/2020GL087253.

——, V. A. Gensini, and B. S. Barrett, 2022: Madden-Julian oscillation influences United States springtime tornado and hail frequency. *npj Climate Atmos. Sci.*, **5**, 37, https://doi.org/10.1038/s41612-022-00263-5.

Mo, K. C., and D. Plettenmaier, 2020: Prediction of flash droughts over the United States. *J. Hydrometeor.*, **21**, 1793–1810, https://doi.org/10.1175/JHM-D-19-0221.1.

Moore, T. W., and M. P. McGuire, 2020: Tornado-days in the United States by phase of the Madden–Julian oscillation and global wind oscillation. *Climate Dyn.*, **54**, 17–36, https://doi.org/10.1007/s00382-019-04983-y.

——, J. M. St. Clair, and T. A. DeBoer, 2018: An analysis of anomalous winter and spring tornado frequency by phase of the El Niño/Southern Oscillation, the Global Wind Oscillation, and the Madden-Julian Oscillation. *Adv. Meteor.*, **2018**, 1–14, https://doi.org/10.1155/2018/3612567.

Murphy, A. H., and E. S. Epstein, 1989: Skill scores and correlation coefficients in model verification. *Mon. Wea. Rev.*, **117**, 572–582, https://doi.org/10.1175/1520-0493(1989)117<0572:SSACCI>2.0.CO;2.

Saha, S., and Coauthors, 2014: The NCEP Climate Forecast System version 2. *J. Climate*, **27**, 2185–2208, https://doi.org/10.1175/JCLI-D-12-00823.1.

Shah, R. D., and V. Mishra, 2015: Development of an experimental near-real-time drought monitor for India. *J. Hydrometeor.*, **16**, 327–345, https://doi.org/10.1175/JHM-D-14-0041.1.

——, and ——, 2016: Utility of Global Ensemble Forecast System (GEFS) reforecast for medium-range drought prediction in India. *J. Hydrometeor.*, **17**, 1781–1800, https://doi.org/10.1175/JHM-D-15-0050.1.

Sherburn, K. D., and M. D. Parker, 2014: Climatology and ingredients of significant severe convection in high-shear, low-CAPE environments. *Wea. Forecasting*, **29**, 854–877, https://doi.org/10.1175/WAF-D-13-00041.1.

Smith, A. B., and R. W. Katz, 2013: US billion-dollar weather and climate disasters: Data sources, trends, accuracy and biases. *Nat. Hazards*, **67**, 387–410, https://doi.org/10.1007/s11069-013-0566-5.

Talib, A., A. R. Desai, J. Huang, T. J. Griffis, D. E. Reed, and J. Chen, 2021: Evaluation of prediction and forecasting models for evapotranspiration of agricultural lands in the Midwest U.S. *J. Hydrol.*, **600**, 126579, https://doi.org/10.1016/j.jhydrol.2021.126579.

Taszarek, M., J. T. Allen, T. Púčik, K. A. Hoogewind, and H. E. Brooks, 2020: Severe convective storms across Europe and the United States. Part II: ERA5 environments associated with lightning, large hail, severe wind, and tornadoes. *J. Climate*, **33**, 10 263–10 286, https://doi.org/10.1175/JCLI-D-20-0346.1.

Thompson, R. L., R. Edwards, J. A. Hart, K. L. Elmore, and P. Markowski, 2003: Close proximity soundings within supercell environments obtained from the rapid update cycle. *Wea. Forecasting*, **18**, 1243–1261, https://doi.org/10.1175/1520-0434(2003)018<1243:CPSWSE>2.0.CO;2.

——, B. T. Smith, J. S. Grams, A. R. Dean, and C. Broyles, 2012: Convective modes for significant severe thunderstorms in the contiguous United States. Part II: Supercell and QLCS tornado environments. *Wea. Forecasting*, **27**, 1136–1154, https://doi.org/10.1175/WAF-D-11-00116.1.

Tippett, M. K., 2018: Robustness of relations between the MJO and U.S. tornado occurrence. *Mon. Wea. Rev.*, **146**, 3873–3884, https://doi.org/10.1175/MWR-D-18-0207.1.

Tseng, K. C., E. A. Barnes, and E. D. Maloney, 2018: Prediction of the midlatitude response to strong Madden-Julian Oscillation events on S2S time scales. *Geophys. Res. Lett.*, **45**, 463–470, https://doi.org/10.1002/2017GL075734.

van Straaten, C., K. Whan, D. Coumou, B. van den Hurk, and M. Schmeits, 2020: The influence of aggregation and statistical post-processing on the subseasonal predictability of European temperatures. *Quart. J. Roy. Meteor. Soc.*, **146**, 2654–2670, https://doi.org/10.1002/qj.3810.

Verlinden, K. L., and D. R. Bright, 2017: Using the second-generation GEFS reforecasts to predict ceiling, visibility, and aviation flight category. *Wea. Forecasting*, **32**, 1765–1780, https://doi.org/10.1175/WAF-D-16-0211.1.

Wang, H., A. Kumar, A. Diawara, D. Dewitt, and J. Gottschalck, 2021: Dynamical–statistical prediction of week-2 severe weather for the United States. *Wea. Forecasting*, **36**, 109–125, https://doi.org/10.1175/WAF-D-20-0009.1.

Wang, X., and C. H. Bishop, 2003: A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci.*, **60**, 1140–1158, https://doi.org/10.1175/1520-0469(2003)060<1140:ACOBAE>2.0.CO;2.

Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2007: The discrete Brier and ranked probability skill scores. *Mon. Wea. Rev.*, **135**, 118–124, https://doi.org/10.1175/MWR3280.1.

Wheeler, M. C., and H. H. Hendon, 2004: An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. *Mon. Wea. Rev.*, **132**, 1917–1932, https://doi.org/10.1175/1520-0493(2004)132<1917:AARMMI>2.0.CO;2.

Wilks, D. S., 2011: Forecast verification. *Int. Geophys.*, **100**, 301–394, https://doi.org/10.1016/B978-0-12-385022-5.00008-7.

Winters, A. C., 2021: Subseasonal prediction of the state and evolution of the North Pacific jet stream. *J. Geophys. Res. Atmos.*, **126**, e2021JD035094, https://doi.org/10.1029/2021JD035094.

Zhang, C., 2005: Madden-Julian Oscillation. *Rev. Geophys.*, **43**, RG2003, https://doi.org/10.1029/2004RG000158.

Zhang, F., Y. Qiang Sun, L. Magnusson, R. Buizza, S. J. Lin, J. H. Chen, and K. Emanuel, 2019: What is the predictability limit of midlatitude weather? *J. Atmos. Sci.*, **76**, 1077–1091, https://doi.org/10.1175/JAS-D-18-0269.1.

Zhou, X., and Coauthors, 2022: The development of the NCEP Global Ensemble Forecast System version 12. *Wea. Forecasting*, **37**, 1069–1084, https://doi.org/10.1175/WAF-D-21-0112.1.

Zhuang, J., and Coauthors, 2022: pangeo-data/xESMF: v0.7.0. Accessed 15 April 2023, https://doi.org/10.5281/ZENODO.7447707.

Brought to you by NOAA Library | Unauthenticated | Downloaded 04/01/25 07:02 PM UTC