Forecasting U.S. Tornado Outbreak Activity and Associated Environments in the Global Ensemble Forecast System (GEFS)

Kelsey Malloy[a] and Michael K. Tippett[a]

[a] Department of Applied Physics and Applied Mathematics, Columbia University, New York, New York

(Manuscript received 23 July 2024, in final form 11 November 2024, accepted 22 January 2025)

ABSTRACT: Tornado outbreaks are high-impact events, often causing significant loss of life and property. This study evaluated the forecast skill of U.S. tornado outbreak activity using the Global Ensemble Forecast System (GEFS), version 12, at lead times up to 2 weeks. Tornado outbreak activity is represented in GEFS using an outbreak index, which relates the likelihood of outbreak-level tornadoes to GEFS forecasts of convective precipitation (CP), storm-relative helicity (SRH), and convective available potential energy (CAPE). GEFS forecasts of the outbreak index are verified against smoothed Storm Prediction Center report data. Since the performance of the outbreak index depends on how well GEFS predicts the index constituents, we also evaluated the climatology and forecast skill of CP, SRH, and CAPE, as well as their covariability. We found that GEFS has a systematic low-CAPE bias and that the forecast skill of the outbreak index is most limited by GEFS forecast skill of CP. We corrected the low-CAPE bias and index seasonality errors via a seasonally and regionally dependent scaling of CAPE and the index, which improved the seasonal cycle and forecast skill of CAPE and tornado outbreak activity in GEFS. Overall, on average, GEFS has its highest forecast skill of tornado outbreak activity during winter and spring—in some cases, positive skill extends beyond week 1 forecast leads—and has its lowest forecast skill during summer.

SIGNIFICANCE STATEMENT: Tornado outbreaks—when multiple tornadoes occur over a short time span—are one of the most extreme forms of severe convective storms. Individual tornadoes are typically regarded as unpredictable except at very short lead times, but broad tornado activity or likelihood might be predictable past the weather time scale (a week or more in advance). We evaluated the forecast skill of U.S. tornado outbreak activity as well as the forecast skill of environmental conditions relevant to the favorability of tornado outbreaks in a state-of-the-art forecast model. We found that one environmental condition that describes storm instability or "fuel"—convective available potential energy—is often too low in the forecast model. We also found that tornado outbreak activity as represented by model environmental conditions has seasonal cycle errors. After correcting for these issues, we determined that the forecast skill of tornado outbreak activity is highest in winter–spring and is lowest in summer. Overall, winter, spring, and fall forecasts of U.S.-wide tornado outbreak activity are skillful past the weather time scale.

KEYWORDS: Severe storms; Storm environments; Tornadoes; Forecast verification/skill; Model evaluation/performance

1. Introduction

The National Oceanic and Atmospheric Administration (NOAA) Storm Prediction Center (SPC) issues areal outlooks for severe convective storm (SCS) activity up to 8 days in advance. These outlooks include probability maps of damaging winds, hail, and tornadoes on days 1 and 2, and severe weather on days 3–8. Short-lead (days 1–3) forecasts of SCS have been steadily improving (Hitchens and Brooks 2012, 2014; Herman et al. 2018), likely due to increases in numerical weather model resolution. Long-lead forecasts (i.e., past day 8 leads) of SCS are much less skillful, though there have been increasing efforts to document and understand the predictability limits of SCS hazards (Gensini and Tippett 2019; Wang et al. 2021). A particular challenge facing tornado forecasters is that numerical weather prediction models do not resolve tornadoes. Storm-scale parameters from high-resolution regional convection-allowing models inform the forecasting process at the relatively short lead times at which they are available (Gallo et al. 2016).

At longer leads, an "ingredients"-based approach based on the output of relatively coarse global models has been effective in providing forecast guidance. The idea is that the likelihood of severe thunderstorm occurrence, especially supercells—the type of strong thunderstorm that most commonly produces impactful hail and tornadoes—is related to the favorability of local environmental conditions (Brooks et al. 1994, 2003). For instance, Tsonevsky et al. (2018) showed that models could predict the environments associated with SCS, especially extremes, out to 7 days on average. In particular, composite parameters, such as the supercell composite parameter (SCP) and the significant tornado parameter (STP), combine information from dynamic and thermodynamic environmental ingredients to measure how supportive atmospheric conditions are for supercell production (Thompson et al. 2003). SCP is the product of convective

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/WAF-D-24-0138.s1.

Malloy's current affiliation: Department of Geography and Spatial Sciences, University of Delaware, Newark, Delaware.

Corresponding author: Kelsey Malloy, kmmalloy@udel.edu

DOI: 10.1175/WAF-D-24-0138.1

available potential energy (CAPE), 0–3-km storm-relative helicity (SRH), and 0–6-km bulk wind shear, normalized so that an SCP value greater than 1 indicates severe hazards are expected. Carbin et al. (2016) and Gensini and Tippett (2019) used SCP to understand long-lead prediction skill of thunderstorm extent. Using SCP, Global Forecast Ensemble System (GEFS) forecasts for tornado events during spring 2016–17 were found to be skillful for up to 9 days in advance (Gensini and Tippett 2019). Wang et al. (2021) used a dynamical–statistical approach for week 2 prediction of tornado and hail activity with SCP as a predictor, and they considered the empirical relationship between GEFS hindcast SCP and observed report frequencies to "calibrate" the prediction. Though SCP is a practical proxy for tornado activity, its units do not directly correspond to probabilities like those in SPC outlooks, and many studies introduce a somewhat arbitrary SCP threshold in order to indicate events and nonevents and evaluate forecast skill. Hill et al. (2020, 2023) used random forests to develop probabilistic forecasts of severe weather hazards, analogous to SPC outlooks, and showed that their model is skillful up to 5-day lead times. Similarly, Battaglioli et al. (2023) used an Additive Regressive Convective Hazard Model (AR-CHaMo; Rädler et al. 2018) to develop probabilistic forecasts of hail and lightning and assessed its forecast skill. These studies did not consider forecasts past day 8, which would correspond to long-range prediction skill, nor delineate between tornadoes and other SCS hazards. Tornado outbreaks—when multiple tornadoes occur over a short time span—represent the most extreme impact from SCS and account for ~80% of tornado-related fatalities (Fuhrmann et al. 2014). It would be valuable to know the extent to which tornado outbreak events specifically can be predicted, in addition to general SCS or tornado activity. We aim to address this question.

The tornado outbreak index from Malloy and Tippett (2024) provides a spatially resolved probability of outbreak-level tornado occurrence given reanalysis environments of convective precipitation, 0–3-km SRH, and CAPE. The index has been useful for understanding the climate modulation of tornado and tornado outbreak activity, such as from ENSO and weather regimes (Tippett et al. 2024), and in detecting upward trends in U.S. tornado outbreak activity (Malloy and Tippett 2024). In addition, the units of the index (probability) correspond well with SPC outlooks (%), and the index matches the general behavior of observed reports. Although the index probabilities are reliable on average, they have systematic regional and seasonal biases, such as being generally too low in winter and generally too high in summer. The effectiveness of the outbreak index as a proxy for tornado outbreak occurrence in forecast applications largely depends on whether its constituents—convective precipitation, SRH, and CAPE—can be skillfully predicted (Murugavel et al. 2012; Rädler et al. 2018, 2019; Koch et al. 2021; Malloy and Tippett 2024). A comprehensive evaluation of the long-lead prediction skill of environments and their covariability and how that relates to the prediction limits of tornado (outbreak) activity is still needed.

The main objective of this study is to document the extent to which tornado outbreak activity can be forecast using the tornado outbreak index from Malloy and Tippett (2024) applied to GEFS output. In addition, we aim to correct the systematic biases in the index climatology to make it more useful in forecast applications. Since the ability to predict the index depends on the ability to predict the environments used to calculate the index, a secondary objective of the study is to document the prediction skill of the environments and their covariability in GEFS. The remainder of the paper is outlined as follows: section 2 describes the data and methods used in the study. Section 3 presents the results in regard to GEFS climatology and calibration, GEFS skill evaluation of tornado outbreak activity, GEFS skill evaluation of relevant environments and covariability of environments, and a GEFS example use case for the spring 2013 season. Section 4 gives a summary of the work and discusses its main conclusions and future work needed.

2. Data and methods

a. Data

Observations of environmental variables of convective precipitation (CP), 0–3-km SRH, and mixed-layer CAPE for the period 2000–19 are taken from the North American Regional Reanalysis (NARR), which provides 3-hourly data on a 32-km native grid. We resample the NARR data as a 6-hourly sum for CP and a 6-hourly average for SRH and CAPE, and we perform a bilinear interpolation of variables to a 1° × 1° spatial resolution.

Tornado reports for the 2000–19 period are taken from the NOAA SPC Severe Weather Database. A tornado outbreak is defined as a sequence of six or more tornadoes that occurs with no more than 6 h between consecutive tornadoes, following the definition used in Fuhrmann et al. (2014) and Malloy and Tippett (2024). We exclude tornadoes rated 0 on the Fujita/enhanced Fujita scale. When labeling outbreak-level tornadoes, we include only the tornado start location and do not impose a geographic constraint (Doswell et al. 2006). Tornadoes are classified as being part of an outbreak or not. The SPC report data are then aggregated to a 6-hourly and 1° × 1° resolution; and at this point, the data are zeros and ones. A one means that an outbreak tornado occurred in a given grid cell and 6-hourly period. For this study, we consider spatially smoothed occurrence data. We apply a 2D Gaussian kernel smoother with $\sigma = 120$ km. This calculation is similar to that used for practically perfect hindcasts (Hitchens et al. 2013; Gensini et al. 2020; Sobash et al. 2020). Smoothing results in values between 0 and 1, consistent with the index. However, the smoothed reports have limitations; for instance, the contribution of adjacent grid points could mean that the grid point of maximum probability might not correspond to where an outbreak tornado occurred.

Model data are taken from the GEFS, version 12, reforecasts (Guan et al. 2022). GEFS reforecasts were initialized once per day at 0000 UTC over the 2000–19 period. GEFS reforecasts have one control member and four perturbed members, equaling five ensemble members in total, that are run out to 16 days with forecast outputs every 6 h. The forecasts initialized on Wednesdays have an additional six ensemble members (11 total members) and run out to 35 days. GEFS data are provided at a 1° × 1° spatial resolution, consistent with observations.

The tornado outbreak index of Malloy and Tippett (2024) has two parts. The first part provides a map of the probability of outbreak tornado occurrence based on

$$\log\left(\frac{p}{1-p}\right) = -20.2 + 0.76\log(\mathrm{CP}) + 1.82\log(\mathrm{SRH}) + 0.51\log(\mathrm{CAPE}), \qquad (1)$$

where the left-hand side is the log odds and p is the probability. The NARR-based index inputs the 6-hourly values of CP, SRH, and CAPE from NARR into the right-hand side, whereas the GEFS-based index inputs the 6-hourly values of CP, SRH, and CAPE from individual GEFS ensemble members. We resample this to a daily resolution by taking the maximum over the convective day (1200–1200 UTC). For GEFS, this means skipping the +6-h time step and taking the maximum over the +12-, +18-, +24-, and +30-h time steps as the day 1 forecast, taking the maximum over the +36-, +42-, +48-, and +54-h time steps as the day 2 forecast, and so on. Finally, we take the ensemble mean of the index, which is used for evaluating performance.

The second part of the index from Malloy and Tippett (2024) calculates the total number of U.S. outbreak tornadoes using the probability maps from above:

$$\mu = \exp\{-0.82 + 2.08\log[\mathrm{sum}(P_{\mathrm{CONUS}})] - 0.57\log[\max(P_{\mathrm{CONUS}})]\}, \qquad (2)$$

where $\mathrm{sum}(P_{\mathrm{CONUS}})$ is the index map sum and $\max(P_{\mathrm{CONUS}})$ is the index map maximum. We compare the index sum, index maximum, and total number of outbreak tornadoes between reports and GEFS forecasts to assess the skill of CONUS-wide tornado outbreak activity.

The seasonal cycle of the environments and index is evaluated. If observed and GEFS seasonal cycles do not match well, we perform a bias correction of the environments and a postcalibration of the index. The procedure consists of multiplying the environments or index by a scaling factor that is the ratio of the seasonally averaged (December–February, March–May, June–August, September–November) observed value to the seasonally averaged GEFS value at every grid point and forecast lead time. This method only corrects for climatological biases and has no dependence on skill. The bias correction of the environments uses NARR data, and the postcalibration of the index uses smoothed reports. We use the term postcalibration since it corrects the GEFS index after its calculation. The index postcalibration is cross validated with 1 year left out, i.e., the postcalibration factor for the 2000 GEFS index is based on 2001–19 data, the postcalibration factor for the 2001 GEFS index is based on 2000 and 2002–19 data, and so on.

b. *Performance metrics*

The mean-square error skill score (MSESS) is used to evaluate how well the GEFS ensemble mean forecast index matches the spatially smoothed report data. The MSESS is the ratio of the mean-square error (MSE) of a forecast to the MSE of a reference forecast:

$$\mathrm{MSESS} = 1 - \frac{\mathrm{MSE}}{\mathrm{MSE}_{\mathrm{ref}}} = 1 - \frac{\sum_i (f_i - o_i)^2}{\sum_i (o_i - \overline{o})^2}, \qquad (3)$$

where f_i is the forecast GEFS ensemble mean probability at time i; o_i is the probability calculated from the spatially smoothed report data; and the reference forecast here is the climatological forecast, where $\overline{o}$ is the time-smoothed (15-day window) daily climatological frequency from the reports. A positive MSESS indicates a forecast skill better than a climatological forecast. As in Wang et al. (2021), we spatially coarsen the GEFS index (f_i) by averaging the data over a 5° × 5° area for the MSESS calculation.

The Spearman rank correlation is also used to measure the strength of the relationship between GEFS forecasts and observed smoothed reports. Both the MSESS and rank correlations are grouped by season, and gridpoint scores are averaged over U.S. regions of interest.

Last, we show reliability diagrams, which measure the conditional bias of GEFS or the extent to which the GEFS ensemble mean index probability matches the observed frequency from the smoothed report data. This usage is slightly different than traditional reliability diagrams in that the observational occurrence data are smoothed and have values between zero and one.

c. *Statistical significance*

We perform 10-day block bootstrap resampling with replacement for 100 iterations to calculate the statistical significance of the seasonal, regionally averaged MSESS and rank correlations. First, the initialization dates are split into 10-day blocks (no overlap); the 10-day blocks account for temporal autocorrelation of the data. For 100 iterations, we resample 10-day blocks (within the corresponding season), select the observed and GEFS forecast data for the resampled dates, and recalculate the regionally averaged MSESS or rank correlation. This procedure constructs a confidence interval for the skill score estimate. If the 5th percentile of bootstrap resamples of MSESS or rank correlation is greater than zero, the positive skill score is considered statistically significant at the 5% level (one-sided). In other words, we consider a positive skill score estimate statistically significant if 95% of the skill scores from the block bootstrap iterations are also positive.

3. Results

a. *Seasonal cycle and index calibration*

We compare the seasonal cycles from GEFS forecasts to observations (reports as well as NARR reanalysis) to determine if there are systematic biases in GEFS. In Fig. 1, the regional seasonal cycles of tornado outbreak probability, environments, and covariability of environments are shown for the northern plains and Midwest, southern plains, Southeast United States, and Northeast United States (see regional domains in Fig. 2). These regional domains are consistent with Hill et al. (2020, 2023) but further subdivide the central plains and eastern United States

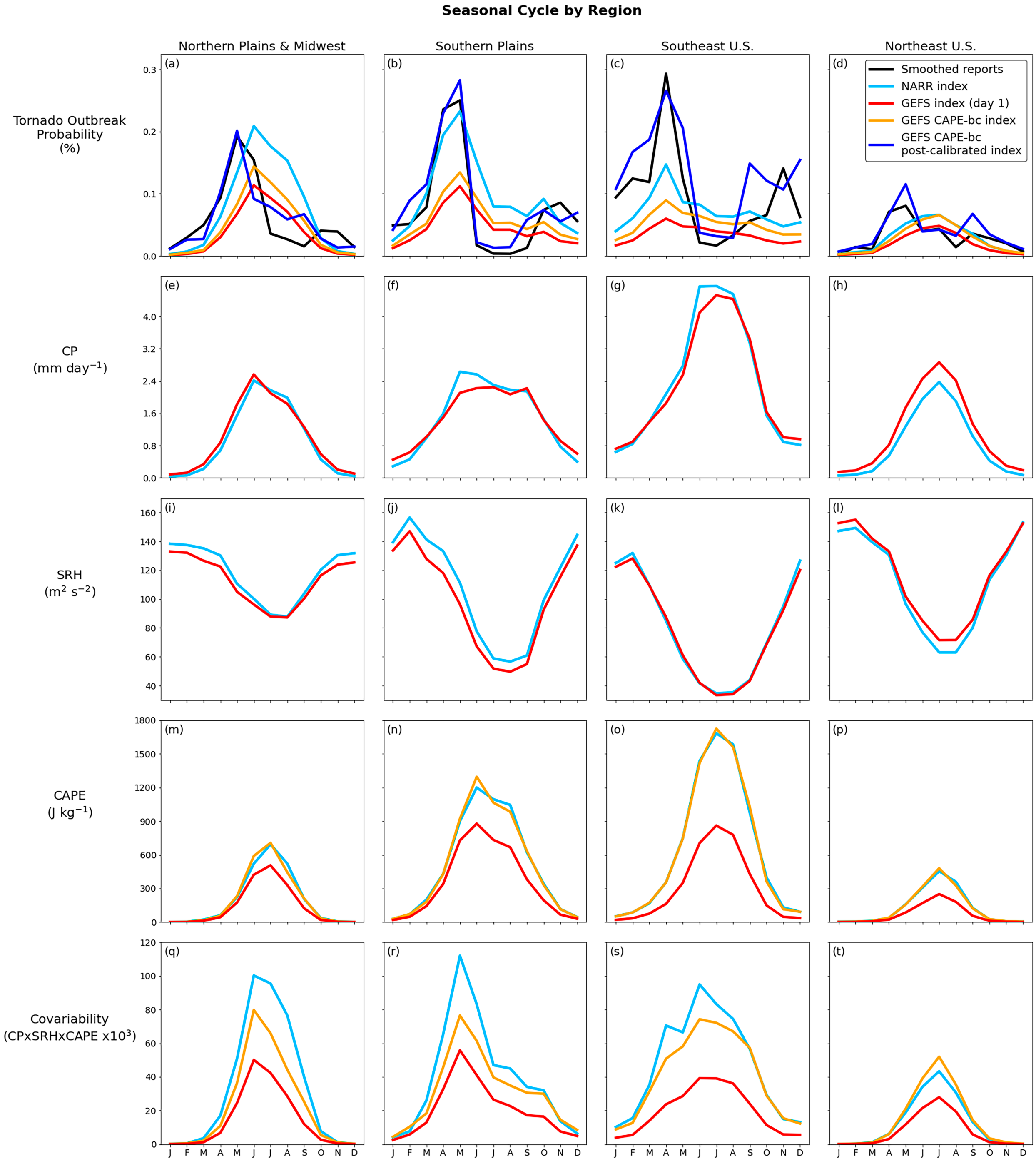

FIG. 1. Seasonal cycle of (a)–(d) tornado outbreak probability, (e)–(h) CP, (i)–(l) SRH, (m)–(p) CAPE, and (q)–(t) covariability (i.e., the gridpoint product) of CP, SRH, and CAPE, for different regions of the United States: (first column) northern plains and Midwest, (second column) southern plains, (third column) Southeast United States, and (fourth column) Northeast United States. Line colors indicate from smoothed reports (black), NARR (light blue), GEFS day 1 forecast (red), GEFS day 1 forecast with bias-corrected CAPE (orange), and GEFS day 1 forecast with bias-corrected CAPE and postcalibration (dark blue). See text for details about CAPE bias correction and postcalibration.

to better qualitatively describe regional patterns in tornado outbreak activity.

First, we evaluate the outbreak index seasonal cycle by comparing the smoothed reports (black line) to the NARR outbreak index (light blue line) in Figs. 1a–d. The comparison between the NARR-based index and smoothed reports identifies fundamental limitations of the index, which is to say, the ability of environments to predict tornado outbreak

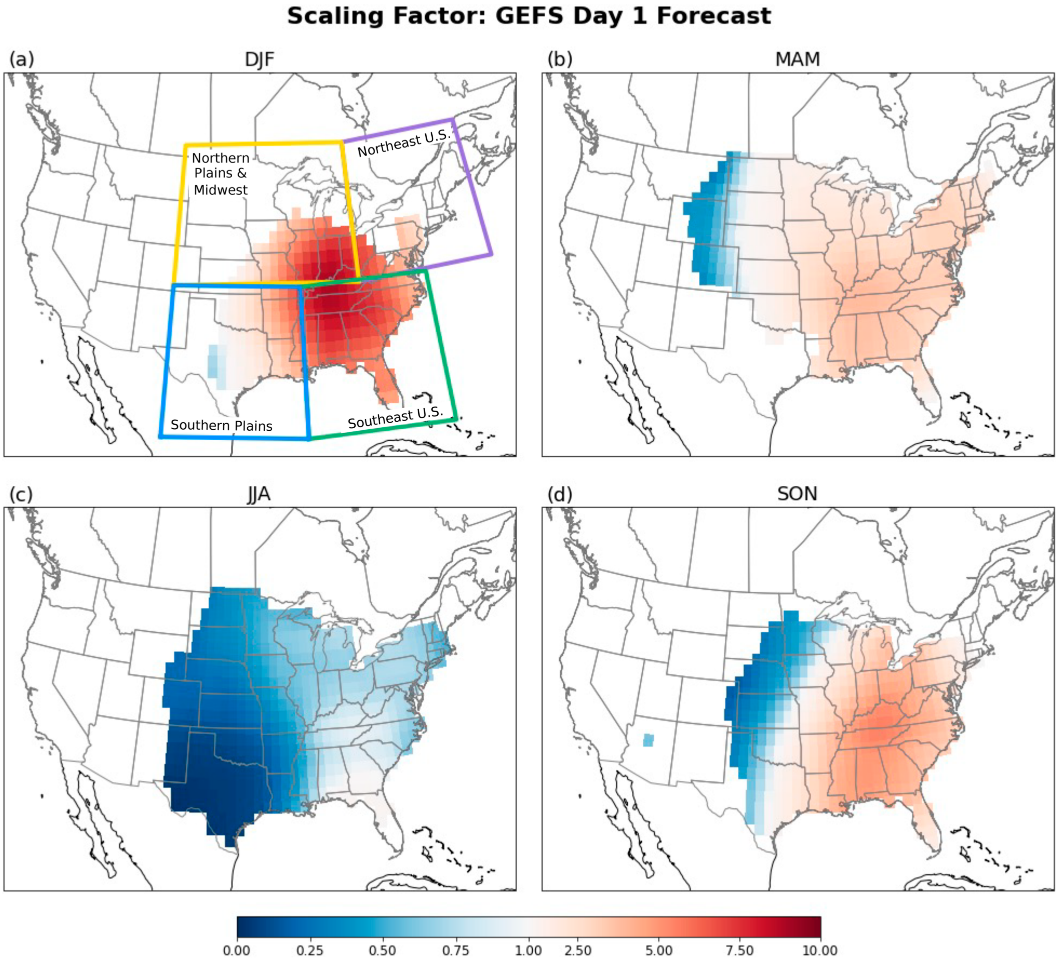

FIG. 2. Scaling factor for GEFS outbreak index day 1 forecasts fit to smoothed reports for (a) DJF, (b) MAM, (c) JJA, and (d) SON. Domain boxes in (a) describe regions used for regionally averaged calculations. Grid points are masked where the tornado outbreak climatological frequency is <0.01%.

occurrence. The NARR-based index is expected to be an upper bound of the performance of the GEFS-based forecast index. As discussed in Malloy and Tippett (2024), the index represents many aspects of the observed seasonal cycle, such as the increase in outbreak-level tornado probability from winter into spring for all four regions. However, the index is too high in the summer months, and the slight peak in activity observed in November is generally not well represented except perhaps over the southern plains. This highlights deficiencies in the index that are expected to also manifest in the GEFS outbreak index forecasts.

Next, we compare the GEFS outbreak index day 1 forecasts (red line) to the NARR outbreak index (light blue line) to identify deficiencies in the GEFS environments that are inputs to the index. The seasonal cycle follows the same pattern in the GEFS index versus NARR index. Yet, overall, the GEFS index is lower—about half the magnitude of the NARR index—in almost all seasons and regions. This suggests that a systematic bias might be present in one or more environments in GEFS.

Turning to the GEFS environments, we compare the seasonal cycles of CP (Figs. 1e–h), SRH (Figs. 1i–l), CAPE (Figs. 1m–p), and the covariability (i.e., gridpoint product) of these variables (Figs. 1q–t) to assess any biases in the environmental conditions. GEFS shows a low-CAPE bias over all regions, especially over the Southeast United States in the summer months (red vs light blue lines). Therefore, we test whether correcting the systematic CAPE biases in GEFS improves its seasonal cycle and, as a result, the GEFS index.

To bias correct the GEFS CAPE, we multiply by a scaling factor that is the ratio of seasonally averaged NARR CAPE to seasonally averaged GEFS CAPE at every grid point and 6-h forecast lead time. This multiplicative bias correction, as opposed to an additive one, has the advantage of allowing low-CAPE values (i.e., zero or near-zero values). Importantly, this correction is based only on climatological values; there is no dependence on skill. The scaling factor for the GEFS CAPE 6–12-h forecast is shown in Fig. S1 in the online supplemental material. CAPE was corrected over regions where tornado outbreaks occur frequently; generally, the multiplication factors range between 4 and 16, depending on the region and season.

After bias correction, the seasonal cycle of GEFS CAPE matches that of NARR well (orange lines in Figs. 1m–p). In addition, the covariability of the environments is improved in GEFS by fixing the systematic biases in CAPE (Figs. 1q–t, orange line), especially over the Southeast U.S. and Northeast U.S. regions. We recalculate the GEFS outbreak index with the bias-corrected CAPE (orange lines in Figs. 1a–d). We find that correcting the systematic biases in CAPE moves the GEFS index closer to the NARR index, especially over the Southeast U.S. and Northeast U.S. regions, but the CAPE bias correction has minimal impact on the GEFS index for the northern plains and Midwest as well as the southern plains.

Comparing the smoothed reports (black line) and GEFS outbreak index day 1 forecasts, raw/without any calibration (red line) or with bias-corrected CAPE (orange line) shows that there are still deficiencies in the seasonal cycle similar to those of the NARR outbreak index.

We address the deficiencies in the index climatology by cross-validated postcalibration, as introduced in the data and methods: applying a scaling factor that is the ratio of the seasonally averaged smoothed reports and the seasonally averaged GEFS index (with bias-corrected CAPE) at every grid point and daily forecast lead time. We also smooth the scaling factor spatially (5° × 5° window) and in terms of lead time (10-day window); this limits the impact from specific reported events in this relatively short 20-yr period. In Fig. 2, we show the maps of the scaling factor for GEFS *day 1* forecasts for December–February (DJF), March–May (MAM), June–August (JJA), and September–November (SON). During DJF, the GEFS index for day 1 forecasts is generally too low over most regions east of the Rockies, as indicated by the scaling factors greater than one. For instance, over the Ohio River Valley, the GEFS index values need to be multiplied by a factor of ~8–10 to better fit the smoothed report data climatology. In contrast, during JJA, the GEFS index for day 1 forecasts is generally too high, especially over the plains. For instance, over Texas, the GEFS index values need to be multiplied by ~0–0.25 to better fit the smoothed report data climatology.

We find that postcalibration of the GEFS index addresses the issues with its seasonal cycle (Figs. 1a–d, dark blue line), including representing the lower activity during summer and a small secondary peak in activity during autumn. For the remainder of the skill analysis, we will compare the skill of the raw GEFS outbreak index, the GEFS outbreak index with bias-corrected CAPE, and the GEFS outbreak index with bias-corrected CAPE and postcalibration.

b. *Outbreak index skill evaluation*

Figure 3 shows the MSESS of the three GEFS-based tornado outbreak indices. On average, the raw GEFS index (red line) generally has very modest but positive skill for forecast leads out to 6–7 days during DJF, 7–9 days during MAM, 5–7 days during JJA (except no skill over the southern plains), and 6–9 days during SON, with variations depending on the region. There is minimal difference between the MSESS for the raw GEFS index (red line) and the GEFS index with bias-corrected CAPE (orange line). Therefore, correcting the low bias in CAPE has little to no impact on the skill in representing index probability values. The MSESS is generally higher for the postcalibrated GEFS index (blue line) than for the raw GEFS index or the GEFS index with bias-corrected CAPE. This shows that correcting the GEFS index seasonal cycle via postcalibration greatly benefits its skill in representing index probability values. This is likely due to the fact that MSESS, unlike correlation, is sensitive to mean and amplitude errors. On average, the postcalibrated GEFS index generally has positive skill for forecast leads of 8–10 days during DJF, 8–9 days during MAM, 1–5 days during JJA, and 5–8 days during SON. These MSESS values are statistically significant (via the boostrap resampling procedure) for 4–9 days during DJF and MAM. Over the southern plains during JJA (Fig. 3j), the postcalibrated GEFS index has positive skill for forecast leads up to 13 days, which is much improved over the negative skill from the raw GEFS index (not shown for being too negative for this y scale) or the GEFS index with bias-corrected CAPE. However, overall, the MSESS values during JJA are very low compared to other seasons. The largest difference between the raw GEFS index and the postcalibrated GEFS index is during DJF over the northern plains, southern plains, and Northeast United States as well as during MAM over Southeast United States, with the postcalibrated GEFS index providing 1–3 more lead days with positive skill.

Next, we consider the rank correlation between observed and GEFS tornado outbreak activity, seen in Fig. 4. Unlike MSESS, rank correlation is unaffected by the CAPE bias correction and postcalibration because multiplication by a scaling factor does not change the ranking (order) of the data. Therefore, we only show the rank correlation for the postcalibrated GEFS outbreak index here. During JJA, the GEFS index generally does not have robust rank correlations > 0.1 for any forecast leads. In addition, the GEFS index generally does not have robust positive rank correlations for any forecast leads during SON and DJF over the northern plains and the Northeast United States, though the lack of robustness in correlations could partly be explained by the smaller sample sizes from masking grid points with tornado outbreak frequencies $< 0.01\%$ (see masked grid points in Fig. 2). On average, the forecast postcalibrated GEFS index has robust rank correlations > 0.1 for forecast leads of 12–13 days during DJF over the southern plains and the Southeast United States, 10–12 days during MAM, and 8–10 days during SON over the southern plains and the Southeast United States. The rank correlation results generally match the MSESS results, showing that the highest skill is during DJF and MAM over all regions and the lowest skill is during JJA for all regions.

Reliability diagrams for the GEFS day 1–7 forecasts of the outbreak index are shown in Fig. 5. Figure 5a shows the reliability for the full year of forecast data, and Figs. 5b–e show the reliability for forecast data separated by season. A well-calibrated model will have reliability curves on the 1:1 line (labeled "perfect"), and skillful forecasts will have reliability curves that fall within the gray shaded region.

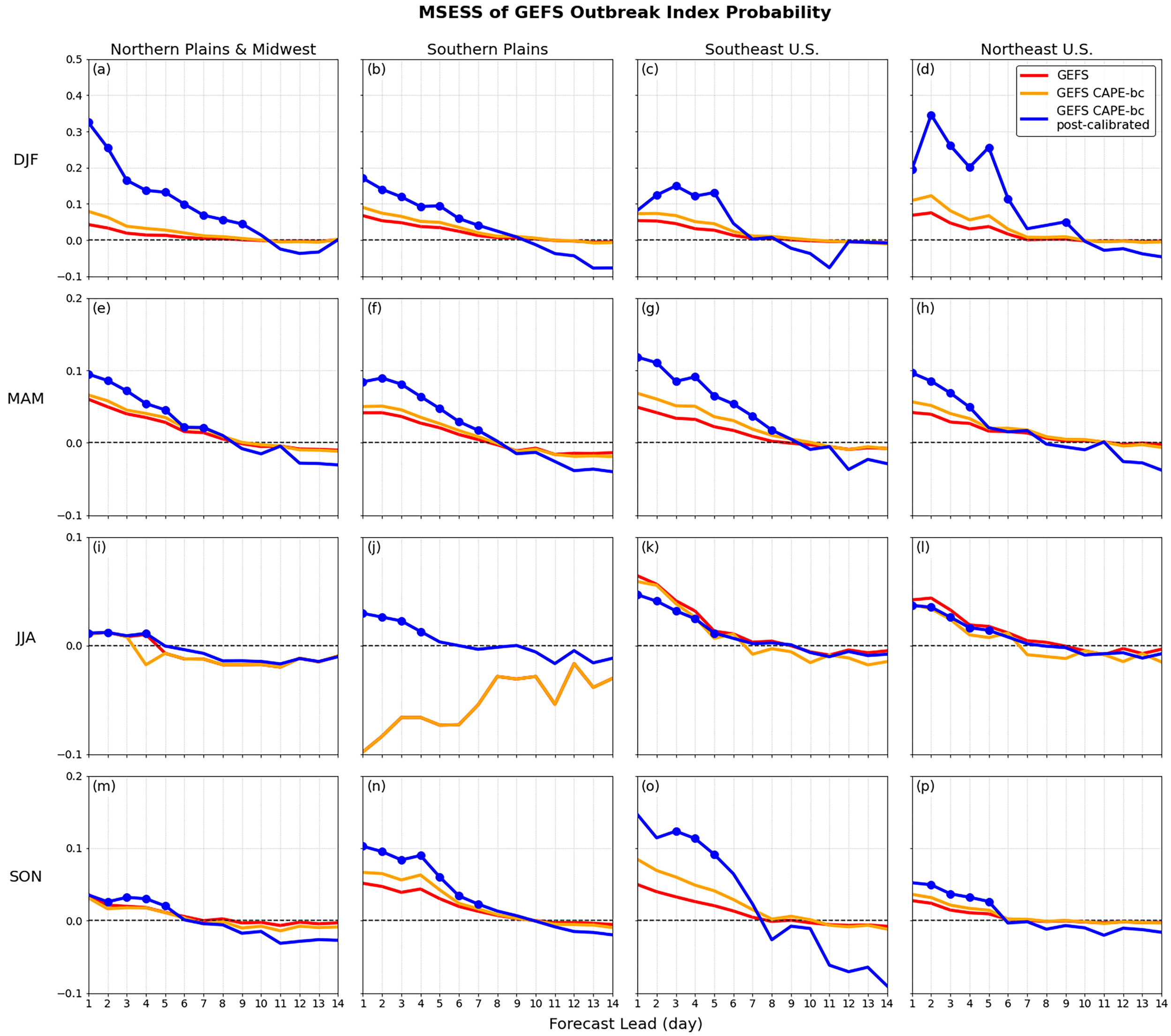

FIG. 3. MSESS for the GEFS outbreak index grouped by (a)–(d) DJF, (e)–(h) MAM, (i)–(l) JJA, and (m)–(p) SON seasons and averaged over (first column) northern plains and Midwest, (second column) southern plains, (third column) Southeast United States, and (fourth column) Northeast U.S. regions. Regionally averaged data exclude grid points where the tornado outbreak climatological frequency is <0.01%. Blue circle markers represent MSESS values from GEFS with bias-corrected CAPE and postcalibration that are greater than 0 and statistically significant at the 5% level.

For the full-year data, the raw GEFS index (red line) is underconfident on average for index probabilities less than 3% (e.g., a GEFS forecast of 2% verifies at 3%), and it is greatly overconfident for index probabilities greater than 6%. The reliability diagrams separated by season reveal that underconfidence for relatively low forecast index probabilities and overconfidence for relatively high forecast index probabilities are found in DJF, MAM, and SON. In addition, JJA forecasts are generally overconfident and unskillful for index probabilities greater than 3%.

The reliability of the postcalibrated GEFS index (blue line) is generally improved over that of the raw GEFS index. For the full year of data, on average, the postcalibrated GEFS index has virtually perfect reliability, i.e., it closely verifies with observed frequency, for index probabilities less than 5%. The postcalibrated GEFS index remains skillful—i.e., it outperforms climatological forecasts—but is overconfident between index probabilities of 5% and 20%. The seasonal reliability diagrams reveal details about the conditional bias of the postcalibrated GEFS index. DJF forecasts have virtually perfect reliability for index probabilities less than 10% and, like with the full year of data, are skillful but overconfident until index probabilities of ~24%. MAM and SON forecasts have virtually perfect reliability for index probabilities less than 5% and, similar to the full year of data, are skillful but overconfident until index probabilities of ~16%; the forecasts are

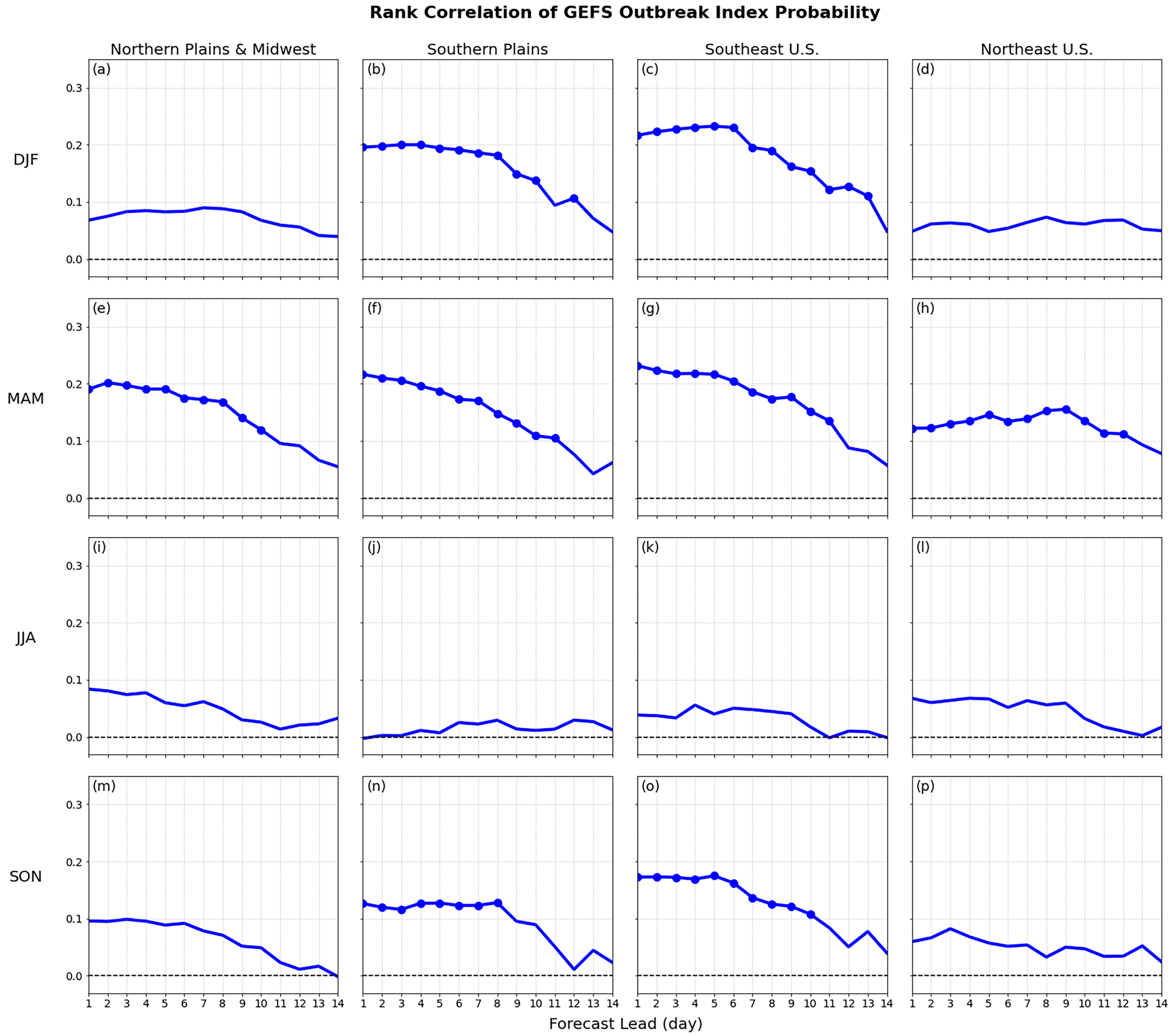

FIG. 4. Rank correlation between reports and the postcalibrated GEFS outbreak index grouped by (a)–(d) DJF, (e)–(h) MAM, (i)–(l) JJA, and (m)–(p) SON seasons and averaged over (first column) northern plains and Midwest, (second column) southern plains, (third column) Southeast United States, and (last column) Northeast U.S. regions. Regionally averaged data ignore grid points where the tornado outbreak climatological frequency is <0.01%. Circle markers represent correlations > 0.1 and statistically significant at the 5% level.

unskillful (do not outperform climatological forecasts) for probabilities greater than 16%. JJA forecasts for the postcalibrated GEFS index do not show improvement over the raw GEFS index for index probabilities greater than 3%; the index is generally overconfident and unskillful for relatively high forecast index probabilities. This behavior is consistent with results from Malloy and Tippett (2024), which found that the outbreak index overforecasts outbreak events during JJA. Interestingly, the postcalibration—which fixes the seasonal cycle of the index—does not appear to improve the reliability (conditional bias) during JJA, perhaps due to the generally low skill noted before.

We also constructed reliability diagrams for GEFS day 8–14 forecasts of the index, which show that the reliability of the index decreases sharply after the first week; forecasts are generally overconfident and cannot outperform climatological forecasts for forecast index probabilities greater than 2% (Fig. S2).

Finally, to understand GEFS forecast skill of CONUS-wide tornado outbreak activity, we examine the skill with which the gridpoint sum and maximum of the daily maps of the smoothed reports, as well as the total number of outbreak tornadoes, can be forecast. Malloy and Tippett (2024) found that the sum and gridpoint maximum value of the index are effective predictors for the total number of U.S. outbreak tornadoes and, therefore, can be used to represent total U.S. tornado outbreak activity. In addition, this removes the requirement that GEFS outbreak activity *spatially* matches observed outbreak activity.

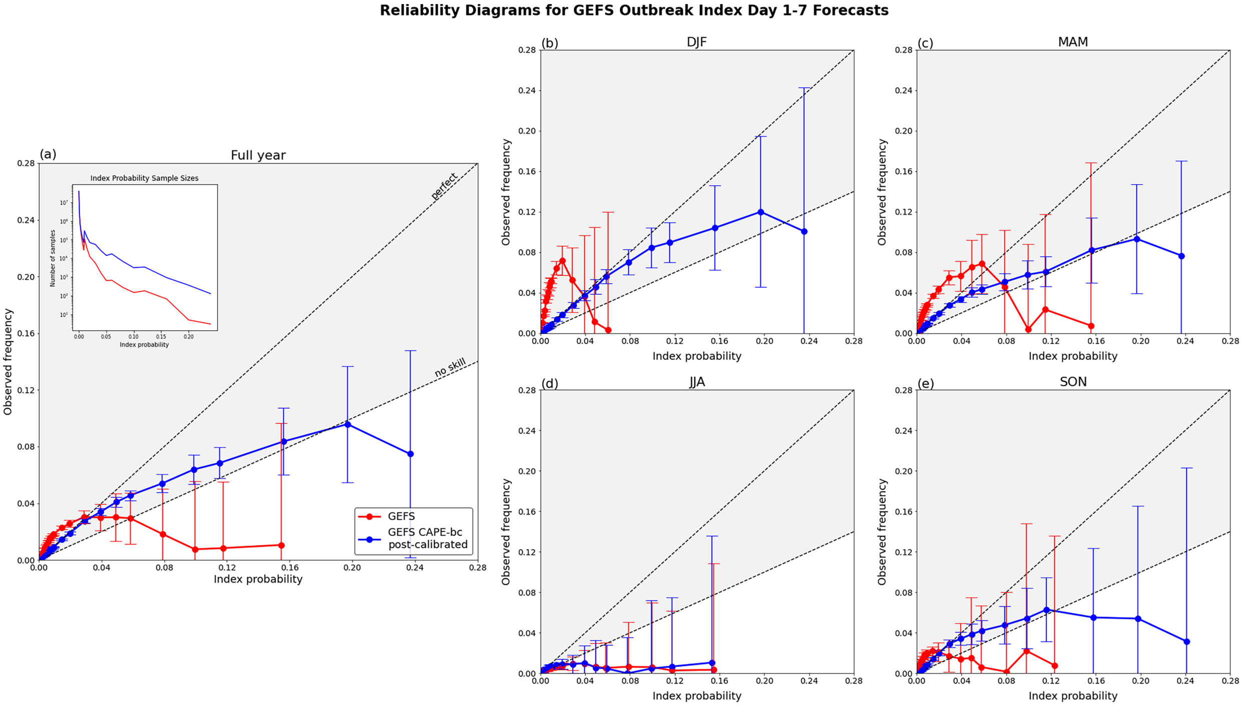

FIG. 5. Reliability diagrams for the raw GEFS outbreak index (red lines) and postcalibrated GEFS outbreak index (blue lines) for (a) the full year of data, as well as data grouped by (b) DJF, (c) MAM, (d) JJA, and (e) SON seasons. The subpanel in (a) shows the sample sizes for each index probability bin for the full year of data. The "perfect" reliability line and "no skill" lines are displayed as dashed lines and are labeled in (a). Light gray shading indicates where forecasts positively contribute to skill scores. Error bars are the 95% confidence interval of the estimate.

Figure 6 shows that the GEFS forecasts of the map sum, maximum, and total number of tornadoes have robust correlations > 0.1 with those of the smoothed reports for lead times out to 13 days in DJF, 11–13 days in MAM, 5–8 days in JJA, and 11 days in SON. Therefore, while there are noted limits in forecasting gridpoint activity on a regional basis, such as for JJA, GEFS forecast skill for CONUS-wide tornado outbreak activity can be skillful for longer forecast leads.

In brief, the GEFS outbreak index has systematic biases related to a low bias of CAPE in GEFS as well as fundamental limitations of the index that result in misrepresentation of the seasonal cycle. We improve the GEFS outbreak index by bias-correcting CAPE and performing a postcalibration of the index, fixing some of these systematic biases in the GEFS outbreak index. Overall, GEFS index skill is highest during DJF and MAM for all regions, perhaps having skill up to days 8–13 on average, and is lowest (with no skill in some cases) during JJA. In addition, GEFS forecast skill in CONUS-wide tornado outbreak activity as represented by the aggregated outbreak index and total number of tornadoes is higher for longer forecast leads. GEFS skill in terms of the *regionally averaged* outbreak index—i.e., the index is averaged over a region before MSESS and rank correlation are computed—is also assessed (Figs. S3 and S4). Results are similar but generally the regionally averaged outbreak index has 1–2 more forecast lead days of positive, robust skill compared to the regionally averaged skill from the index at each grid point.

c. *Environment skill evaluation*

We hypothesize that the limitations in forecast skill for the raw and bias-corrected GEFS outbreak index might be partly explained by the forecast skill of the environments that are used for the calculation of the index—CP, SRH, and CAPE—as well as the forecast skill of their covariability, i.e., the product of convective (CP and CAPE) and kinematic (SRH) variables. Therefore, in the next set of results, we evaluate the skill via rank correlation for CP, SRH, and CAPE as well as the rank correlation of the products CP × SRH, CAPE × SRH, and CP × SRH × CAPE.

The rank correlation for the GEFS environments is shown in Fig. 7. The variable with the lowest skill is CP (blue lines) for almost all seasons and regions. CP forecasts have no robust positive correlation with observed CP for any forecast leads over the northern plains and Midwest during DJF. Otherwise, on average, CP forecasts have robust positive rank correlations with observed CP for forecast leads of 4 days over the Northeast United States and 10–11 days over the southern plains and the Southeast Unites States during DJF, 11–13 days during MAM, 8–9 days during JJA, and 9–11 days during SON. Both forecast SRH (gold lines) and forecast CAPE (green lines) have robust positive correlations with observations for longer forecast leads. On average, SRH forecasts are skillful for forecast leads of 10–12 days during DJF, 11–13 days during MAM, 9–11 days during JJA, and 11–13 days during SON. CAPE forecasts are skillful for forecast leads of 9–13 days

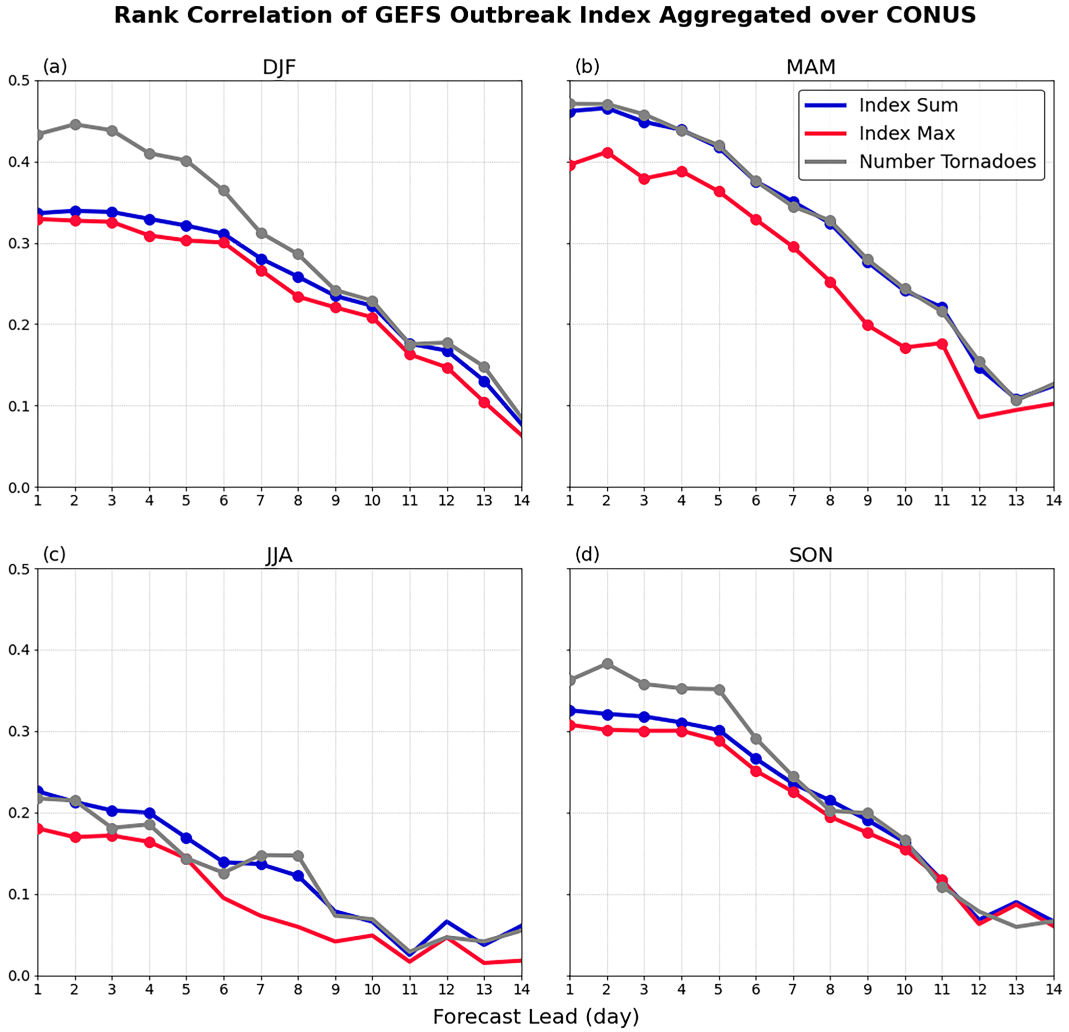

FIG. 6. Rank correlation between the CONUS map sum (red), gridpoint maximum (blue), and number of total tornadoes (gray) from observed smoothed reports and the map sum, gridpoint maximum, and total number of outbreak tornadoes from the postcalibrated GEFS outbreak index grouped by (a) DJF, (b) MAM, (c) JJA, and (d) SON seasons. Circle markers represent correlations that are >0.1 and statistically significant at the 5% level.

during DJF and 12–13 days during MAM, JJA, and SON. This suggests that CP is the convective variable that limits forecast skill for the GEFS index. However, it is also important to understand how well GEFS forecasts the covariability of the environments since tornado outbreak activity is generally conditional on the presence of both convective and kinematic variables.

The rank correlations of the covariability of the environments are shown in Fig. 8. Overall, the correlations for the covariability of the environments are similar to or lower than the correlations for the individual environmental variables. Forecast CP–SRH (orange–red lines) has robust positive correlations with observed CP–SRH for similar lead times as seen with CP forecasts, supporting the idea that the limited ability to forecast CP limits the forecast skill of its covariability. For instance, as with CP forecasts, CP–SRH forecasts have no or modest robust positive correlations over the northern plains and Midwest and Northeast United States. On the other hand, forecast CAPE-SRH (turquoise lines) has the highest correlations with observed CAPE–SRH and are statistically significant and >0.2 for the longest lead times on average. Therefore, GEFS's ability to forecast SRH and CAPE as well as their covariability positively contributes to forecast skill for tornado outbreak activity past the weather time scale. The forecast CP–SRH–CAPE (pink lines) has correlations that fall somewhere between the other correlations in covariability, likely due to competing influences from the relatively poorly forecast CP and the relatively well-forecast SRH and CAPE.

d. Example case

We illustrate the behavior of the index by considering GEFS forecasts of the outbreak index during spring 2013, including how the GEFS outbreak index forecasts evolved for a historical tornado outbreak event, shown in Fig. 9. A chiclet plot is used

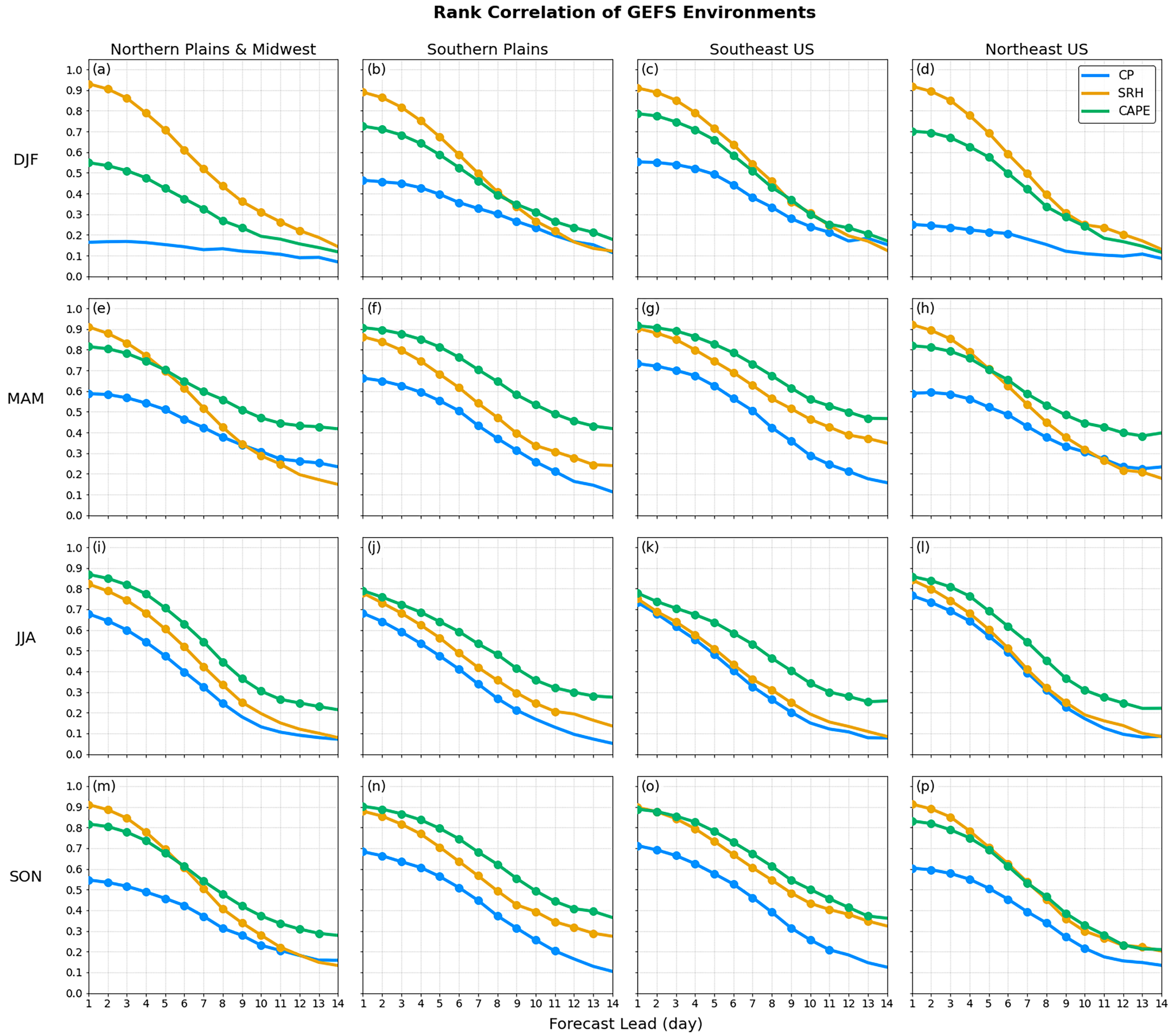

FIG. 7. Rank correlation between NARR and GEFS CP (blue), SRH (gold), and CAPE (green) grouped by (a)–(d) DJF, (e)–(h) MAM, (i)–(l) JJA, and (m)–(p) SON seasons and averaged over (first column) northern plains and Midwest, (second column) southern plains, (third column) Southeast United States, and (fourth column) Northeast U.S. regions. Regionally averaged data ignore grid points where the tornado outbreak climatological frequency is <0.01%. Circle markers represent correlations that are >0.2 and statistically significant at the 5% level.

to visualize the progression of forecasts depending on their valid date (x axis) and lead time (y axis) (Fig. 9a, top half). Here, we are showing forecasts of the anomalous (i.e., deviations from climatology) index sum over CONUS (shaded). Single forecast runs are found on the forward diagonals. When considering a fixed valid date, which corresponds to a single column in the chiclet plot, vertical features indicate consistently forecast (across initializations) episodes of enhanced tornado activity, and their height shows how far in advance they were predicted. For instance, for 10 April 2013, GEFS signals high U.S. tornado activity 11 days in advance; for 17 April, GEFS signals high activity 6 days in advance; for 20 May, GEFS signals high activity 10 days in advance; and for 31 May, GEFS signals high activity 9 days in advance. We can compare this to the total number of outbreak-level reports over CONUS for each date (bottom half). In general, the (valid) dates when GEFS has relatively large positive index sum values for multiple forecast leads correspond to days when outbreak tornadoes were reported. In fact, many outbreak days during this season had a positive index sum in GEFS for forecast leads considered past the weather time scale (>7-day forecast lead). We also note the anomalously low index sum values between late April and mid-May that correspond to an observed "quiet" period in terms of U.S. outbreak activity.

We further analyze GEFS forecasts on a high-impact outbreak event, 20 May 2013, seen in Figs. 9b–f. On 20 May 2013, an upper-level trough moved over the central United States. In the afternoon, moisture, shear, and instability increased

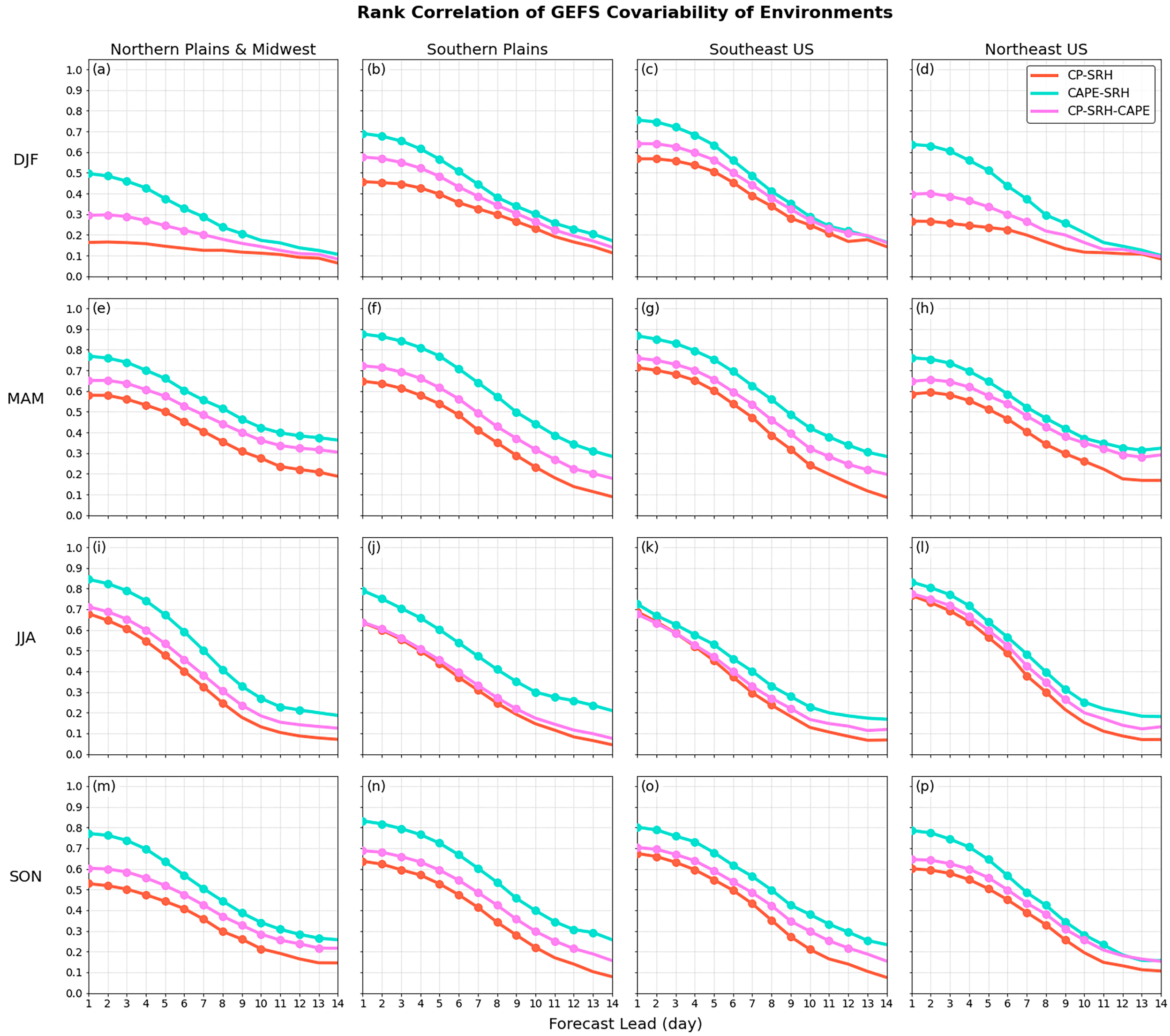

FIG. 8. As in Fig. 7, but for the rank correlation between NARR and GEFS covariability of environments: CP–SRH (orange–red), CAPE–SRH (turquoise), and CP–SRH–CAPE (pink).

rapidly over the central and southern plains, favoring the development of supercells (also see Figs. S5a–d). Many violent tornadoes occurred, including one rated five on the enhanced Fujita (EF) scale, which is the highest level in terms of damage. Twenty-five people were killed, over 300 people were injured, and the event caused $2 billion in damage. In Fig. 9b, the SPC day 1 outlook is shown (contours) as well as the observed May 20 tornado locations and numbers (blue shaded dots). The SPC issued a maximum of 10% probability for tornadoes across the plains. Tornadoes occurred across Texas, Oklahoma, Kansas, Arkansas, Missouri, and Illinois.

Figures 9c and 9d present the GEFS day 1 forecast, and Figs. 9e and 9f present the GEFS day 10 forecast. The GEFS day 1 forecast shows an ensemble mean index with relatively high outbreak tornado probabilities over the general region outlined by the SPC and where most tornadoes were reported, including the highest probabilities over the central plains (Fig. 9c). The ensemble mean index also shows relatively high outbreak tornado probabilities over parts of the Appalachian region, different from the SPC outlook and a region that did not have reports. There is general agreement in the location of high outbreak tornado probabilities, seen by the 100% of ensemble members that agree on these locations experiencing index values greater than the 50th percentile (Fig. 9d). The GEFS 10-day forecast shows an ensemble mean index with high outbreak tornado probabilities over the Tennessee River Valley, slightly more east than the observed event (Fig. 9e). About 60%–80% of ensemble members agree with this general region of outbreak tornado probabilities being greater than the 50th percentile (Fig. 9f). Overall, GEFS indicates general regions of high tornado outbreak activity for forecast leads past the weather time scale. The longer lead times for

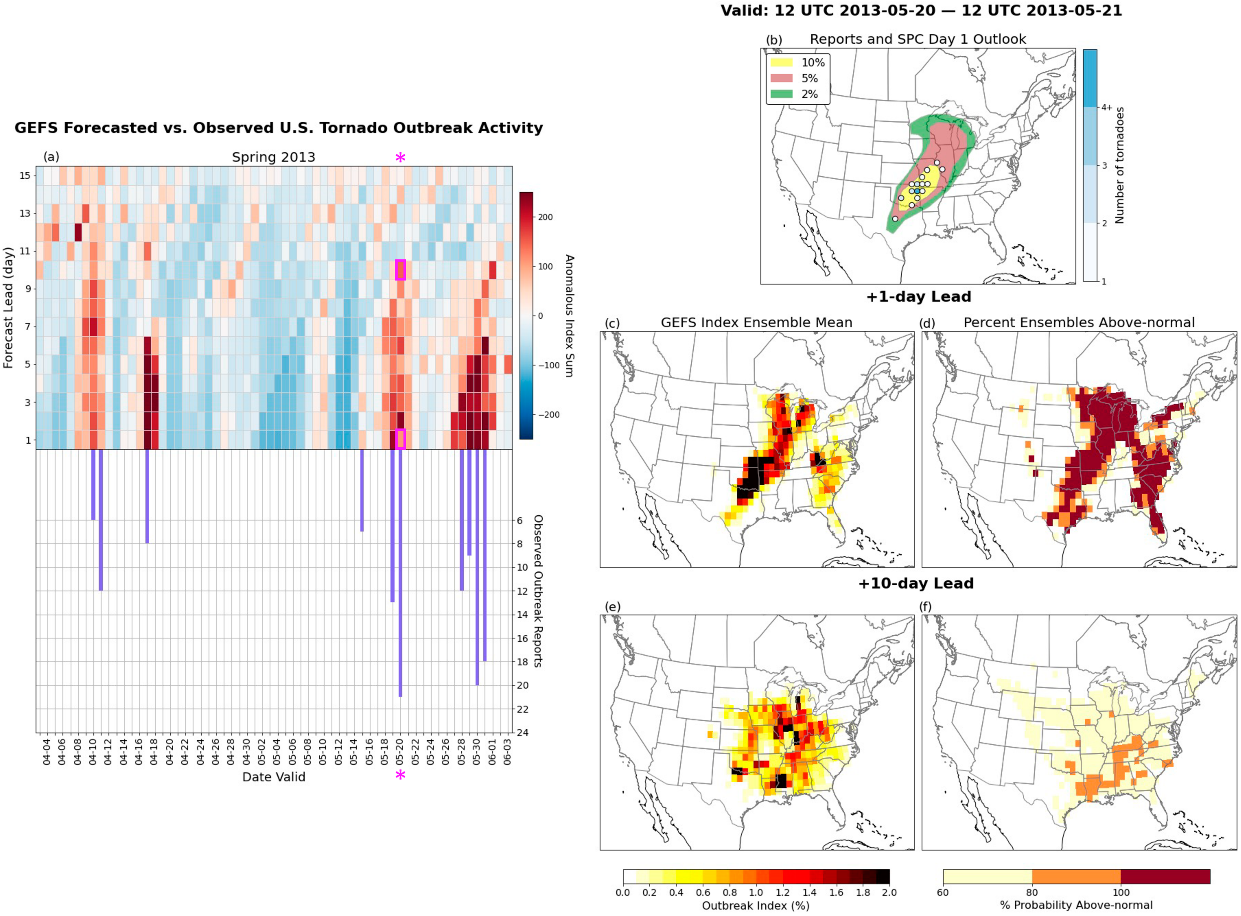

FIG. 9. Example use case for the GEFS outbreak index: (a) April–June 2013 tornado outbreak activity from (top) GEFS forecasts of (shaded) anomalous tornado outbreak index summed over CONUS as a function of valid date and forecast lead, and (bottom) corresponding number of U.S. reports for the valid date; 20 May 2013 is marked with pink asterisks, and day 1 and day 10 forecasts are outlined in solid pink; (b) observed (blue shaded circles) number of tornadoes at grid point and the SPC day 1 outlook at 1200 UTC 20 May 2013–1200 UTC 21 May 2013; GEFS day 1 forecast (initialized at 0000 UTC 20 May) of (c) outbreak index from the ensemble mean and (d) percentage of ensemble members with an above-median (>50th percentile) index; and (e),(f) as in (c) and (d), but for the day 10 forecast (initialized at 0000 UTC 11 May).

this event are likely due to the influence from large-scale, synoptic activity, which can often provide signals in long-lead forecasts. For instance, the GEFS day 10 ensemble mean forecast of 500-hPa geopotential height still indicated a weak trough in the region and associated favorable environments of CP, SRH, and CAPE, which would contribute to an overall above-normal likelihood of tornado outbreak activity (Figs. S5i–l). In contrast, many severe weather events are triggered from mesoscale systems that are considered less predictable for long leads (Carbin et al. 2016; Sobash et al. 2016).

4. Summary and discussion

The purpose of this study is to document the GEFS, version 12, forecast skill of U.S. tornado outbreak activity and relevant environmental factors. To represent tornado activity in GEFS, we calculated a tornado outbreak index from Malloy and Tippett (2024), which models outbreak-level tornado probabilities at every grid point along with an associated number of tornadoes based on collocated values of CP, SRH, and CAPE.

By examining the seasonal cycle of the GEFS outbreak index and corresponding predictors (Fig. 1), we found that GEFS has a systematic low-CAPE bias that results in index biases. These CAPE biases are greater than a factor of four in the afternoon hours, present over most regions and during most times of the year (Fig. S1). In addition, the index has errors in its seasonality—i.e., underestimation of activity during DJF and overestimation of activity during JJA—due to factors not accounted for in the index (i.e., this bias is also present in the NARR-based index). Therefore, first we bias-corrected CAPE in GEFS and recalculated the outbreak index. Then, we applied a postcalibration method to the GEFS outbreak index (Fig. 2). Together, these methods improve the climatological representation of tornado outbreak activity in GEFS.

We showed that the postcalibration benefits GEFS MSESS and reliability in particular (cf. Figs. 3 and 5). Overall, GEFS skill in forecasting tornado outbreak activity is highest during winter and spring, showing low skill into week 2 on average (Figs. 3 and 4). In contrast, GEFS has little to no skill in forecasting the tornado outbreak index during summer. Reliability diagrams revealed that the GEFS index is generally overestimating the relatively high probabilities (>16% for winter, spring, and autumn, or >3% for summer). The forecast underconfidence as seen in the seasonal reliability diagrams (cf. Fig. 5, red lines) is consistent with the broad regions of the postcalibration scaling factor being greater than one in all seasons except JJA (cf. Fig. 2), and overconfidence is consistent with broad regions of the scaling factor being less than one. GEFS has slightly better skill in forecasting CONUS-wide tornado outbreak activity (Fig. 6), or region-wide tornado outbreak activity (cf. Figs. S3 and S4), than tornado outbreak activity at individual grid points, showing some skill into week 2.

To help explain GEFS skill in the outbreak index forecasts, we evaluated GEFS forecast skill in representing the environmental variables and their covariability. In general, the highest skill is shown for SRH and CAPE, and the lowest skill is shown for CP. Furthermore, the covariability of SRH and/or CAPE with CP is lower than the covariability of SRH and CAPE. This suggests that GEFS forecast skill for the outbreak index is most limited by the GEFS forecast skill of CP.

Finally, we demonstrated how the GEFS index forecasts represent the spring 2013 season and a historical tornado outbreak event, 20 May 2013, at a 10-day lead. GEFS forecasts signal increased U.S. tornado outbreak activity for that day, including high index probabilities over the general region where observed reports are located.

This work extends upon previous studies that have estimated medium- to long-range forecast skill of SCS in GEFS (Tsonevsky et al. 2018; Gensini and Tippett 2019; Wang et al. 2021; Hill et al. 2023). By focusing on tornado outbreaks and separating GEFS forecast skill by season and region, we provided a detailed skill assessment of extreme SCS events, which has not been done previously. Tornado outbreak occurrence was modeled by the outbreak index from Malloy and Tippett (2024), which has no explicit dependence on location or season, only dependence on environments. Here, we found that a postcalibration that added seasonal and regional dependence to the outbreak index improved its climatology and usefulness in forecast applications. We found that the 5° × 5° spatial coarsening of the index before calculating MSESS and rank correlation does improve the scores compared to no smoothing (not shown), as in Wang et al. (2021). The sporadic nature of tornadoes means predicting exact locations—even if related to the spatial distribution of CP, SRH, and CAPE—is a challenge. In addition, the poor performance of the GEFS outbreak index in summer is consistent with previous work. Convective modes during summer might differ from convective modes during other seasons (Hart and Cohen 2016; Smith et al. 2012), and therefore, forecasting outbreak activity via this outbreak index is not as feasible during summer.

In addition, by examining the forecast skill of environments and their covariability, we attempt to explain the environmental factors that benefit or limit GEFS forecast skill. While the low bias in CAPE has been noted by operational forecasters as potentially related to planetary boundary layer processes (A. M. Bentley 2024, personal communication; Tallapragada 2022; Sun et al. 2024), this work documents its seasonal cycle biases compared to NARR and shows the impact of this bias on the representation of tornado outbreak activity here and SCS activity in general. We found that these biases did not change substantially with forecast lead time (not shown), suggesting that the biases were due to model deficiencies rather than model initialization/drift issues. Interestingly, although the CAPE seasonal cycle improved with the bias correction, the seasonal cycle of covariability and the index did not improve over all regions, and skill did not improve with the CAPE bias correction. Perhaps CAPE has conditional biases or biases with a different functional dependence, both of which the scaling factor does not address. The low skill in forecasting CP relative to SRH and CAPE might suggest that improving the representation of convective processes in models might lead to improving prediction skill of outbreak activity. Alternatively, an outbreak index that does not include CP could have better predictive skill despite being a worse index when computed using reanalysis data. There might be a trade-off between constructing an index that effectively captures observed tornado outbreak activity and performs well in a forecast context.

Future work should further explore the ensemble predictability of tornado outbreak activity from GEFS. In this study, we use the GEFS ensemble mean tornado outbreak index, not taking advantage of the full set of ensembles to understand the predictability of tornado outbreak activity (cf. Figs. 9d,f). A next step could be forecasting the chance of above- or below-normal activity based on the ensembles, an approach used in other studies (Lee et al. 2021). Another next step would be to test the outbreak index and its postcalibration for GEFS real-time forecasts. Future work should also consider if there are periods of enhanced prediction skill of the index due to ENSO, MJO, or other sources of climate variability. Miller and Gensini (2023) found that GEFS forecasts initialized in warm ENSO and active MJO were more skillful on average. Kim et al. (2024) found that a strongly negative and long-lived Pacific–North American (PNA) pattern preceded many high-impact tornado outbreaks. In general, ENSO can modulate tornado and tornado outbreak frequency and has been linked to SCS predictability (Lepore et al. 2017, 2018; Tippett and Lepore 2021; Tippett et al. 2022; Malloy and Tippett 2024).

Given that there are times when SCS can be forecast 8+ days in advance (cf. Fig. 9, Miller and Gensini 2023), and winter and spring show average forecast skill of tornado outbreaks into week 2, it might be of value to develop guidance products for medium- to long-range severe weather alongside SPC or Climate Prediction Center (CPC) operational forecasters.

Acknowledgments. The authors acknowledge the support of this research by the Willis Research Network (Grant WILLIS CU15-2366) and NOAA (Grant NA19OAR4590159). The authors also would like to thank three anonymous reviewers for their insightful comments.

Data availability statement. NARR data are provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their website at https://psl.noaa.gov/data/gridded/data.narr.html. Storm report observations are provided by NOAA/SPC at https://www.spc.noaa.gov/wcm/#data. NOAA GEFS v12 reforecast data are provided by Amazon Web Services at https://noaa-gefs-retrospective.s3.amazonaws.com/index.html.

REFERENCES

Battaglioli, F., P. Groenemeijer, I. Tsonevsky, and T. Púčik, 2023: Forecasting large hail and lightning using additive logistic regression models and the ECMWF reforecasts. *Nat. Hazards Earth Syst. Sci.*, **23**, 3651–3669, https://doi.org/10.5194/nhess-23-3651-2023.

Brooks, H. E., C. A. Doswell III, and J. Cooper, 1994: On the environments of tornadic and nontornadic mesocyclones. *Wea. Forecasting*, **9**, 606–618, https://doi.org/10.1175/1520-0434(1994)009<0606:OTEOTA>2.0.CO;2.

——, J. W. Lee, and J. P. Craven, 2003: The spatial distribution of severe thunderstorm and tornado environments from global reanalysis data. *Atmos. Res.*, **67–68**, 73–94, https://doi.org/10.1016/S0169-8095(03)00045-0.

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.

Doswell, C. A., III, R. Edwards, R. L. Thompson, J. A. Hart, and K. C. Crosbie, 2006: A simple and flexible method for ranking severe weather events. *Wea. Forecasting*, **21**, 939–951, https://doi.org/10.1175/WAF959.1.

Fuhrmann, C. M., C. E. Konrad, M. M. Kovach, J. T. McLeod, W. G. Schmitz, and P. G. Dixon, 2014: Ranking of tornado outbreaks across the United States and their climatological characteristics. *Wea. Forecasting*, **29**, 684–701, https://doi.org/10.1175/WAF-D-13-00128.1.

Gallo, B. T., A. J. Clark, and S. R. Dembek, 2016: Forecasting tornadoes using convection-permitting ensembles. *Wea. Forecasting*, **31**, 273–295, https://doi.org/10.1175/WAF-D-15-0134.1.

Gensini, V. A., 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.

——, A. M. Haberlie, and P. T. Marsh, 2020: Practically perfect hindcasts of severe convective storms. *Bull. Amer. Meteor. Soc.*, **101**, E1259–E1278, https://doi.org/10.1175/BAMS-D-19-0321.1.

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.

Hart, J. A., and A. E. Cohen, 2016: The challenge of forecasting significant tornadoes from June to October using convective parameters. *Wea. Forecasting*, **31**, 2075–2084, https://doi.org/10.1175/WAF-D-16-0005.1.

Herman, G. R., 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, 2012: Evaluation of the Storm Prediction Center's day 1 convective outlooks. *Wea. Forecasting*, **27**, 1580–1585, https://doi.org/10.1175/WAF-D-12-00061.1.

——, and ——, 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.

——, ——, and M. P. Kay, 2013: Objective limits on forecasting skill of rare events. *Wea. Forecasting*, **28**, 525–534, https://doi.org/10.1175/WAF-D-12-00113.1.

Kim, D., S.-K. Lee, H. Lopez, J.-H. Jeong, and J.-S. Hong, 2024: An unusually prolonged Pacific-North American pattern promoted the 2021 winter Quad-State Tornado Outbreaks. *npj Climate Atmos. Sci.*, **7**, 133, https://doi.org/10.1038/s41612-024-00688-0.

Koch, E., J. Koh, A. C. Davison, C. Lepore, and M. K. Tippett, 2021: Trends in the extremes of environments associated with severe U.S. thunderstorms. *J. Climate*, **34**, 1259–1272, https://doi.org/10.1175/JCLI-D-19-0826.1.

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, 2017: ENSO-based probabilistic forecasts of March–May U.S. tornado and hail activity. *Geophys. Res. Lett.*, **44**, 9093–9101, https://doi.org/10.1002/2017GL074781.

——, ——, and ——, 2018: CFSv2 monthly forecasts of tornado and hail activity. *Wea. Forecasting*, **33**, 1283–1297, https://doi.org/10.1175/WAF-D-18-0054.1.

Malloy, K., and M. K. Tippett, 2024: A stochastic statistical model for U.S. outbreak-level tornado occurrence based on the large-scale environment. *Mon. Wea. Rev.*, **152**, 1141–1161, https://doi.org/10.1175/MWR-D-23-0219.1.

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.

Murugavel, P., S. D. Pawar, and V. Gopalakrishnan, 2012: Trends of Convective Available Potential Energy over the Indian region and its effect on rainfall. *Int. J. Climatol.*, **32**, 1362–1372, https://doi.org/10.1002/joc.2359.

Rädler, A. T., P. Groenemeijer, E. Faust, and R. Sausen, 2018: Detecting severe weather trends using an Additive Regressive Convective Hazard Model (AR-CHaMo). *J. Appl. Meteor. Climatol.*, **57**, 569–587, https://doi.org/10.1175/JAMC-D-17-0132.1.

——, P. H. Groenemeijer, E. Faust, R. Sausen, and T. Púčik, 2019: Frequency of severe thunderstorms across Europe expected to increase in the 21st century due to rising instability. *npj Climate Atmos. Sci.*, **2**, 30, https://doi.org/10.1038/s41612-019-0083-7.

Smith, B. T., R. L. Thompson, J. S. Grams, C. Broyles, and H. E. Brooks, 2012: Convective modes for significant severe thunderstorms in the contiguous United States. Part I: Storm classification and climatology. *Wea. Forecasting*, **27**, 1114–1135, https://doi.org/10.1175/WAF-D-11-00115.1.

Sobash, R. A., C. S. Schwartz, G. S. Romine, K. R. Fossell, and M. L. Weisman, 2016: Severe weather prediction using storm surrogates from an ensemble forecasting system. *Wea. Forecasting*, **31**, 255–271, https://doi.org/10.1175/WAF-D-15-0138.1.

——, G. S. Romine, and C. S. Schwartz, 2020: A comparison of neural-network and surrogate-severe probabilistic convective hazard guidance derived from a convection-allowing model. *Wea. Forecasting*, **35**, 1981–2000, https://doi.org/10.1175/WAF-D-20-0036.1.

Sun, X., D. Heinzeller, L. Bernardet, L. Pan, W. Li, D. Turner, and J. Brown, 2024: A case study investigating the low summertime CAPE behavior in the Global Forecast System. *Wea. Forecasting*, **39**, 3–17, https://doi.org/10.1175/WAF-D-22-0208.1.

Tallapragada, V., 2022: Implementation of Global Ensemble Forecast System (GEFSv12) as the first UFS medium range and sub-seasonal weather application. UFS Webinar Series, 82 pp., https://www.ufs.epic.noaa.gov/wp-content/uploads/2020/06/Tallapragada_UFS_Webinar_GEFS-v12_052120.pdf.

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.

Tippett, M. K., and C. Lepore, 2021: ENSO-based predictability of a regional severe thunderstorm index. *Geophys. Res. Lett.*, **48**, e2021GL094907, https://doi.org/10.1029/2021GL094907.

——, ——, and M. L. L'Heureux, 2022: Predictability of a tornado environment index from El Niño–Southern Oscillation (ENSO) and the Arctic Oscillation. *Wea. Climate Dyn.*, **3**, 1063–1075, https://doi.org/10.5194/wcd-3-1063-2022.

——, K. Malloy, and S. H. Lee, 2024: Modulation of U.S. tornado activity by year-round North American weather regimes. *Mon. Wea. Rev.*, **152**, 2189–2202, https://doi.org/10.1175/MWR-D-24-0016.1.

Tsonevsky, I., C. A. Doswell III, and H. E. Brooks, 2018: Early warnings of severe convection using the ECMWF extreme forecast index. *Wea. Forecasting*, **33**, 857–871, https://doi.org/10.1175/WAF-D-18-0030.1.

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.