A Seasonal Probabilistic Outlook for Tornadoes (SPOTter) in the Contiguous United States Based on the Leading Patterns of Large-Scale Atmospheric Anomalies

SANG-KI LEE,[a] HOSMAY LOPEZ,[a] DONGMIN KIM,[a,b] ANDREW T. WITTENBERG,[c] AND ARUN KUMAR[d]

[a] *NOAA/Atlantic Oceanographic and Meteorological Laboratory, Miami, Florida*
[b] *Cooperative Institute for Marine and Atmospheric Studies, University of Miami, Miami, Florida*
[c] *NOAA/Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey*
[d] *NOAA/Climate Prediction Center, College Park, Maryland*

(Manuscript received 14 July 2020, in final form 6 November 2020)

ABSTRACT: This study presents an experimental model for Seasonal Probabilistic Outlook for Tornadoes (SPOTter) in the contiguous United States for March, April, and May and evaluates its forecast skill. This forecast model uses the leading empirical orthogonal function modes of regional variability in tornadic environmental parameters (i.e., low-level vertical wind shear and convective available potential energy), derived from the NCEP Coupled Forecast System, version 2, as the primary predictors. A multiple linear regression is applied to the predicted modes of tornadic environmental parameters to estimate U.S. tornado activity, which is presented as the probability for above-, near-, and below-normal categories. The initial forecast is carried out in late February for March–April U.S. tornado activity and then is updated in late March for April–May activity. A series of reforecast skill tests, including the jackknife cross-validation test, shows that the probabilistic reforecast is overall skillful for predicting the above- and below-normal area-averaged activity in the contiguous United States for the target months of both March–April and April–May. The forecast model also successfully reforecasts the 2011 super-tornado-outbreak season and the other three most active U.S. tornado seasons in 1982, 1991, and 2008, and thus it may be suitable for an operational use for predicting future active and inactive U.S. tornado seasons. However, additional tests show that the regional reforecast is skillful for March–April activity only in the Ohio Valley and South and for April–May activity only in the Southeast and Upper Midwest. These and other limitations of the current model, along with the future advances needed to improve the U.S. regional-scale tornado forecast, are discussed.

KEYWORDS: Tornadoes; Convective storms; Climate prediction; Forecast verification/skill; Probability forecasts/models/distribution; Seasonal forecasting

1. Introduction

During the spring months of March–May (MAM), the central United States east of the Rocky Mountains is most prone to severe thunderstorms that often spawn off a series of violent tornadoes, causing casualties and property losses. For instance, during 2009–18, tornadoes in the United States claimed 890 lives and caused $20 billion in property and crop damages (https://www.spc.noaa.gov/wcm/). Current operational forecasts for severe thunderstorm and tornado hazards (e.g., convective outlooks) are issued a few days in advance. Yet there is a pressing need for expanding severe weather outlooks beyond the synoptic weather time scale toward subseasonal-to-seasonal time scales, to provide emergency managers, government officials, businesses, insurers, and the public advance warning of the potential for loss of life and damage to critical infrastructure.

Previous studies, especially during the past few years, have advanced our understanding of the large-scale atmosphere, ocean and sea ice environments conducive to U.S. tornado outbreaks (e.g., Marzban and Schaefer 2001; Brooks et al. 2003; Marsh et al. 2007; Cook and Schaefer 2008; Muñoz and Enfield 2011; Tippett et al. 2012; Weaver et al. 2012; Barrett and Gensini 2013; Lee et al. 2013, 2016; Thompson and Roundy 2013; Elsner and Widen 2014; Allen et al. 2015, 2018; Saide et al. 2015; Jung and Kirtman 2016; Molina et al. 2016, 2018; Cook et al. 2017; Lepore et al. 2017, 2018; Baggett et al. 2018; Childs et al. 2018; Trapp and Hoogewind 2018; Chu et al. 2019; Molina and Allen 2019). For example, Brooks et al. (2003) derived the low-level vertical wind shear (WSHR) and convective available potential energy (CAPE) threshold values leading to tornadic environmental conditions in the United States. Using similar criteria, Tippett et al. (2012) reasonably reproduced the number of U.S. tornadoes during 1971–2010. Lee et al. (2013) showed that seven of the ten most severe tornado outbreaks in the United States during 1950–2010 were linked to a positive-phase Trans-Niño condition (i.e., colder sea surface temperature anomalies in the central equatorial Pacific than in the eastern equatorial Pacific), which often occurs during the decay phase of La Niña in boreal spring. Allen et al. (2015) showed that La Niña events persisting into boreal spring are linked to increased tornado activity in the central United States, while El Niño events persisting into boreal spring are linked to decreased tornado activity in the central United States. Jung and Kirtman (2016) and Molina et al. (2016) stressed the moisture supply from the Gulf of Mexico as a critical factor that modulates tornado activity in the southern United States. Lee et al. (2016) showed that U.S. regional patterns of tornado outbreak risk are linked to the four main flavors of El Niño–Southern Oscillation (ENSO) in boreal spring, and to tripole variations in North Atlantic sea surface temperature. Lepore et al. (2017, 2018) showed that

Corresponding author: Dr. Sang-Ki Lee, sang-ki.lee@noaa.gov

DOI: 10.1175/MWR-D-20-0223.1

ENSO could modulate monthly and seasonal forecast skill of U.S. tornado activity during MAM, with higher skill during La Niña conditions. Most recently, Trapp and Hoogewind (2018) showed that Arctic sea ice loss may weaken the midlatitude zonal winds and vertical wind shear over North America, suppressing U.S. tornado activity in summer.

These and other recent studies have collectively shown that U.S. tornado activity is directly linked to large-scale regional tornadic environmental parameters (e.g., WSHR and CAPE), which are modulated by ENSO and other slowly varying ocean and sea ice processes. Building upon these findings, here we present and test a hybrid statistical-dynamical seasonal forecast model for U.S. tornado activity. This forecast model uses WSHR and CAPE derived from the NCEP Coupled Forecast System, version 2 (CFSv2; Saha et al. 2014), as the primary predictors, with the premise that the modulating impacts of ENSO and other slowly varying ocean and sea ice processes are integrated into these two tornadic environmental parameters. A multiple linear regression analysis is then applied to the predicted WSHR and CAPE to estimate the likelihood of above-, near-, or below-normal U.S. tornado activity in MAM.

The study is organized as follows. Section 2 presents the tornado index, the atmospheric reanalysis, and the CFSv2 forecast data used in this study. Section 3 analyzes the variability and predictability of WSHR and CAPE using the atmospheric reanalysis and CFSv2. Sections 4 and 5 evaluate the probabilistic reforecast skill of the hybrid statistical-dynamical model for tornadic risk during the period of 1982–2018. The skill evaluation is carried out separately for the contiguous United States (CONUS) and each of the four U.S. climate regions vulnerable to tornadoes (i.e., Ohio Valley, South, Southeast, and Upper Midwest). The probabilistic reforecast for U.S. regional-scale tornado activity (i.e., tornado activity normalized for each of $1° \times 1°$ grid point over the United States) is also presented and its skill is evaluated. Sections 6 and 7 conclude the study with a discussion and summary.

2. Data and methods

a. Tornado data, atmospheric reanalysis, and CFSv2

Several datasets are used to develop and evaluate the seasonal forecast model. We use the Severe Weather Database (SWD) of the National Oceanic and Atmospheric Administration (NOAA), available for downloading (https://www.spc.noaa.gov/wcm/), to identify EF1–EF5 tornadoes in the United States during MAM from 1982 to 2018. Note that EF0 tornadoes are excluded to avoid a spurious long-term trend in the SWD (e.g., Verbout et al. 2006; Lee et al. 2013). To represent the area-averaged tornado activity for the contiguous United States and each of the four climate regions, EF1 tornadoes are also excluded to focus on high-impact tornadoes. However, EF1 tornadoes are included for representing regional-scale tornado activity in $1° \times 1°$ grid boxes over the contiguous United States, to increase the sample size for reliable statistical analysis. To avoid multicounting, the location and EF-scale of each tornado are determined at the time when each tornado achieves its maximum EF-scale (Lee et al. 2016). The European Centre for Medium-Range Weather Forecasts–Interim (ERA-Interim) reanalysis for the period of 1979–2018 is used to derive WSHR (850–1000 hPa) and CAPE anomalies (Dee et al. 2011). CFSv2 reforecasts (1982–2011) and operational forecasts (2011–18) are used as the primary dynamic component of the hybrid statistical-dynamical forecast model. We use 20-member ensemble forecasts initialized every fifth day of February, and four cycles (i.e., 0, 6, 12, and 18 h) from each day, for the target months of March–April (MA). Note that May is excluded from the target months, because a preliminary analysis indicates very little predictability of WSHR and CAPE anomalies for May in the February-initialized CFSv2 forecasts. Therefore, here we use an alternative strategy to update the seasonal forecast for the target months of April–May (AM) using 20-member ensemble forecasts initialized in every fifth day of March.

b. Tornado days versus numbers

To build a seasonal forecast model for tornadoes, we first need to identify an appropriate tornado predictand. Both tornado days and tornado numbers are widely used to represent U.S. tornado activity (e.g., Verbout et al. 2006). Tornado days are computed by counting the number of days in a given period exceeding a threshold number of tornadoes. Tornado numbers are simply the number of tornadoes for a given period. Figure 1 shows these two tornado indices based on EF2–EF5 tornadoes in March–April (MA) and AM during 1954–2018. Also shown are the numbers of tornado-related fatalities in MA and AM during the same periods. All time series are normalized separately for the reforecast period (1982–2018) and the earlier period (1954–81). The correlations between the two indices are statistically significant at the 99% level, based on a Student's t test. However, while the numbers of tornado-related fatalities are highly correlated with tornado numbers, they are poorly correlated with tornado days. The common practice of applying linear correlation (i.e., Pearson correlation) in this case may be limited by the skewness (i.e., non-Gaussian distribution) of the tornado indices. An alternative is to use rank correlation methods, such as Kendall's tau and Spearman's rho, that replace the tornado indices in each year to their rankings among the 65-yr time series, and thus are less sensitive to extreme years like 2011. Spearman's rho method is applied to find that the results from rank correlation analysis are largely consistent.

Many of the historical tornado outbreak seasons are dominated by small number of extreme convective days. For instance, MAM in 1974 is a historical outbreak season in term of tornado numbers, but is a near-normal season in terms of tornado days. On the other hands, MAM in 2011 is an outbreak season in terms of both tornado numbers and days (Fig. 1). Interestingly, the seasonal tornadic environmental parameters were highly favorable in both the 1974 and 2011 seasons (Lee et al. 2013). Therefore, it is still debatable whether the large-scale background tornadic environmental parameters are linked more closely to tornado numbers or tornado days. Nevertheless, since the goal of this seasonal forecast model is to provide advance warning of highly active seasons like the 1974 and 2011 seasons, it is preferable to use tornado numbers as the forecast target. Therefore, the numbers of EF2–EF5

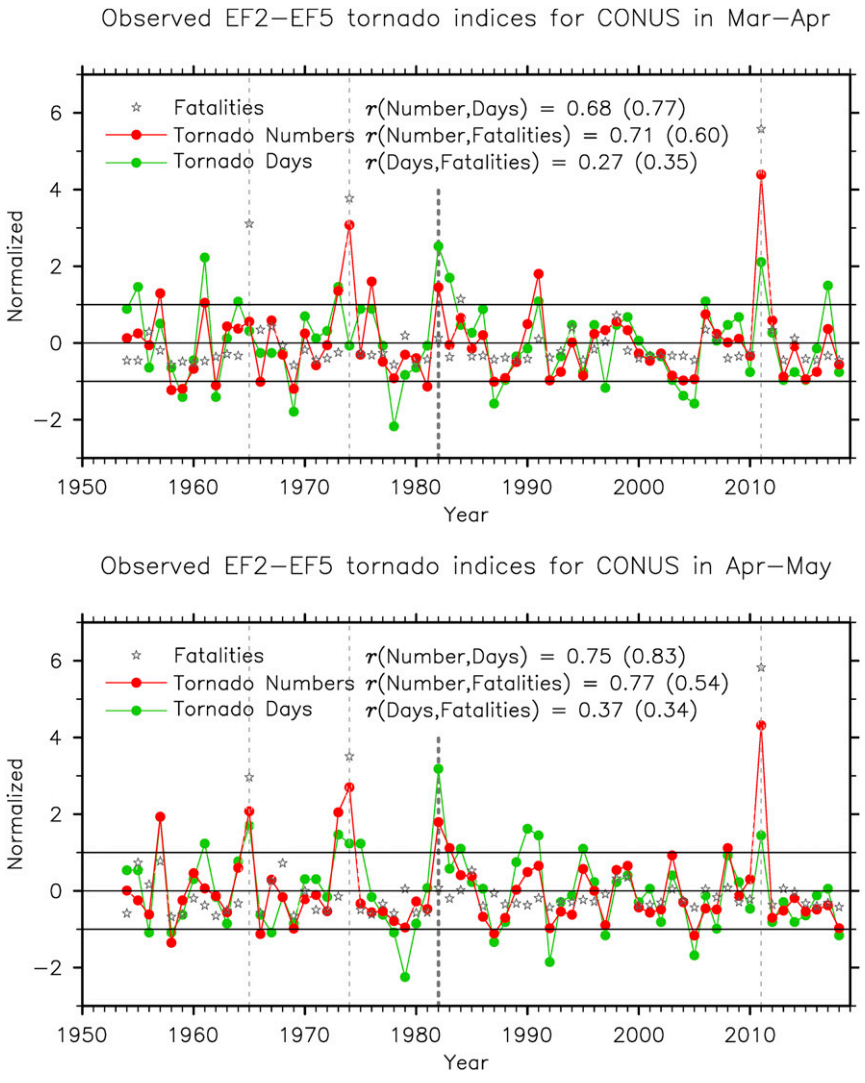

FIG. 1. Tornado numbers and tornado days for the contiguous United States during 1954–2018, as based on EF2–EF5 tornadoes in (top) MA and (bottom) AM, respectively, as diagnosed from SWD. The corresponding numbers of tornado-related fatalities for the same period in MA and AM are also shown. The tornado days are computed by counting the number of days in MA and AM of each year with at least one EF2–EF5 tornado. The tornado numbers are the number of EF2–EF5 tornadoes during MA and AM of each year. Each of these time series is normalized (i.e., the mean is removed, and the result is then divided by its standard deviation) separately for the reforecast period (1982–2018) and the earlier period (1954–81). Also shown are the correlations between the two tornado indices, between the tornado numbers and the number of tornado-related fatalities, and between the tornado days and the number of tornado-related fatalities. The values in parentheses are rank correlations from the Spearman's rho method. The thin vertical gray dashed lines indicate the three deadliest historical U.S. tornado seasons in 1965 (287 fatalities in MAM), 1974 (328 fatalities in MAM), and 2011 (542 fatalities in MAM). The thick vertical gray dashed line indicates the starting year (i.e., 1982) of the reforecast analysis.

tornadoes during MA and AM are used in this study to represent the area-averaged tornado activity in the contiguous United States and each of the four climate regions. To represent regional-scale tornado activity in the United States, we also use a tornado density index, which represents the numbers of EF1–EF5 tornadoes that occur within a 200-km radius of each $1° \times 1°$ grid point over the contiguous United States during MA and AM (Lee et al. 2016).

c. *Evaluation of probabilistic forecast skill*

To evaluate the probabilistic forecast skill of the model, we use the ranked probabilistic skill score (RPSS) and relative

TABLE 1. Contingency table for the probabilistic forecast verification.

	Forecasts		
Observations	Warnings W	Nonwarnings W'	Total
Event E	H	M	e
Nonevent E'	F	C	e'
Total	w	w'	n

operating characteristic (ROC) curve and score along with more traditional statistical tools. RPSS is a skill score that compares the cumulative squared probability error [i.e., the ranked probability score (RPS)] of the probabilistic forecasts for all three categories (i.e., above-, near-, and below-normal activity) with the RPS of the climatology (i.e., 33% chance for each of the three categories). It ranges from 1.0 (i.e., the perfect skill score) to negative infinity. An RPSS value above 0 indicates that the probabilistic forecast is better than the climatological forecast, whereas a value below 0 indicates that the probabilistic forecast skill is worse than that of the climatological forecast.

ROC is a method to evaluate the probabilistic skill of a prediction system based on a 2×2 contingency table (e.g., Swets 1973; Mason and Graham 1999). For example, Table 1 is a contingency table for a system with n observations of e events and e' nonevents, for which w warnings and w' nonwarnings were forecast. For each forecast there are four possible outcomes: a hit if a warning is issued and the event occurs; a miss if no warning is issued for an event that occurs; a false alarm if a warning is issued and no event occurs; and a correct rejection if no warning is issued and no event occurs. The total numbers of hits, misses, false alarms, and correct rejections are given by h, m, f, and c, respectively. The probabilistic skill of a prediction system can be evaluated by comparing hit rates with false alarm rates (e.g., Swets 1973; Mason and Graham 1999). The hit rate (HR) is the proportion of events for which warnings were issued correctly; it provides an estimate of the probability that an event is correctly predicted. The false alarm rate (FAR) is the proportion of nonevents for which warnings were issued incorrectly. These ratios can be written as

$$\text{hit rate} = \frac{h}{h+m} = \frac{h}{e} = p(W|E) \quad \text{and} \tag{1}$$

$$\text{false alarm rate} = \frac{f}{f+c} = \frac{f}{e'} = p(W|E'). \tag{2}$$

For each of the three categories (i.e., above-, near-, and below-normal activity), the contingency table can be constructed and further used to plot the ROC curve, which compares hit rates and false alarm rates for a range of warning threshold values. The ROC curve can be used to find the optimal warning threshold, an application-dependent best tradeoff between hit rate and false alarm rate, for each of the three categories. The ROC score is the area under the ROC curve. It ranges between 0 and 1, and measures the utility of the forecasts compared to the utility of a perfect forecast. An ROC score of 0.5–0.6 generally indicates no forecast skill relative to random guesses from the climatological probability density function, whereas an ROC score above 0.7 indicates that the forecast fairly well discriminates between events and nonevents better than a random guess from the climatological probability density function, so that the system is much more likely to correctly predict an actual event than to issue a false alarm. An ROC score of 0.6–0.7 generally indicates that the forecast skill is poor-to-marginal. Further details on RPSS, the ROC curve, and their applications for meteorological and climate problems can be found in Swets (1973), Harvey et al. (1992), Mason and Graham (1999), Kharin and Zwiers (2003), Hamill and Juras (2006), and Lopez and Kirtman (2014).

3. Variability and predictability of background WSHR and CAPE

As shown in earlier studies, tornadogenesis is closely tied to WSHR and CAPE, and occurs predominantly when both WSHR and CAPE exceed certain threshold values (e.g., Brooks et al. 2003). Therefore, we first use ERA-Interim to explore the leading modes of WSHR and CAPE variability and their linkages to U.S. tornado activity. This is achieved by performing empirical orthogonal function (EOF) analysis of WSHR and CAPE separately for MA and AM over the region most vulnerable to tornado activity (30°–40°N and 100°–80°W).

The independent sets of the first two EOFs of WSHR and CAPE variability are shown in Figs. 2a and 2b for MA and in Figs. 2e and 2f for AM. The two sets of the first EOF (EOF1) of WHSR and CAPE, which explain about 45%–62% of the WSHR and CAPE variances, describe WSHR and CAPE variability over the broad U.S. region east of the Rockies (Figs. 2a,e), and are linked to tornado activity in the South, Ohio Valley, and Southeast (Figs. 2c,g). The two sets of the second EOF (EOF2) of WSHR and CAPE, which account for 15%–29% of the WSHR and CAPE variances, describe dipole-like variability of WSHR and CAPE between the regions northwest and southeast of the Ohio River (Figs. 2b,f), and are linked to dipole-like tornado density variability, with one pole over Oklahoma and Kansas, and the other pole over Arkansas, Mississippi, Tennessee, and Alabama (Figs. 2d,h).

A multiple linear regression analysis, with the independent sets of EOF1 and EOF2 time series of WSHR and CAPE variability (i.e., a total of four time series) as the independent variables (predictors), reasonably well simulates the normalized numbers of EF2–EF5 tornadoes for the contiguous United States, with high correlations between the predicted and observed numbers of tornadoes ($r = 0.68$ for both MA and AM, not shown). This suggests that the independent sets of EOF1 and EOF2 of CAPE and WSHR variability explain more than 45% of the variance in the total number of EF2–EF5 tornadoes for both MA and AM.

The next task is to apply the same EOF analysis to the CFSv2 forecasts, to test if there is any useful skill in predicting the leading EOFs of WSHR and CAPE. Figure 3 is identical to Fig. 2, but derived from the CFSv2 forecasts of WSHR and CAPE. As discussed in section 2a, we use 20 ensemble members initialized in February of each year for the MA forecasts, and those initialized in March for the AM forecasts for the

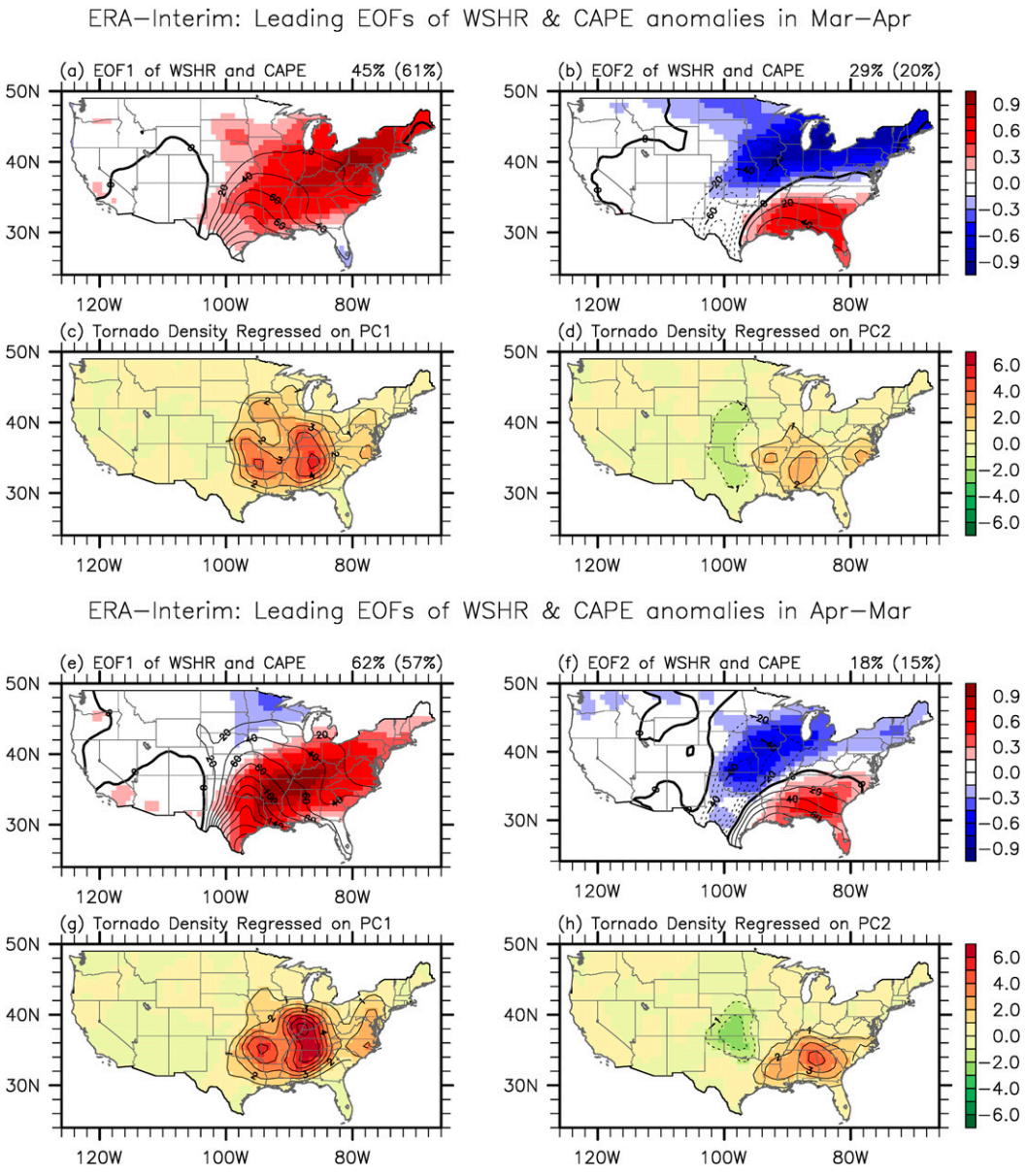

FIG. 2. Major spatial patterns of WSHR (850 − 1000 hPa; shaded) and CAPE (contoured) variability in (a), (b) MA and (e), (f) AM during 1979–2018 derived from ERA-Interim. The WSHR and CAPE fields are regressed onto their corresponding time series of the (left) first and (right) second principal components (PC1 and PC2). The WSHR and CAPE variances explained by each of the two EOFs are also indicated (values in parentheses are for CAPE). Also shown is the observed tornado density field regressed onto the two sets of (c),(g) PC1 and (d),(h) PC2 time series of WSHR and CAPE variability. To compute the regression coefficients in (c) and (g), the two sets of PC1 of WSHR and CAPE variability are combined to perform a multiple linear regression analysis to best fit the tornado density averaged over the box region of 30°–40°N, 100°–80°W. Similarly, the two sets of PC2 of WSHR and CAPE variability are used to best fit the tornado density averaged over the southern U.S. box region (30°–35°N, 100°–80°W), which is then used to compute the regression coefficients in (d) and (h).

period of 1982–2018. All 20 ensemble forecasts are first merged into a single long time series consisting of 740 samples (i.e., 37 years × 20 ensemble members) before performing the EOF analysis. The two sets of EOF1, which explain about 62%–69% of the forecast WSHR and CAPE variances, describe WSHR and CAPE variability over the broad U.S. region east of the Rockies (Figs. 3a,e) consistent with the two sets of EOF1 derived from ERA-Interim (Figs. 2a,e), and are mainly linked to tornado density variability across Arkansas and Mississippi in the South, the Ohio Valley, and Alabama and Georgia in the Southeast (Figs. 3c,g). The two sets of EOF2, which explain about 13%–23% of WHSR and CAPE variances,

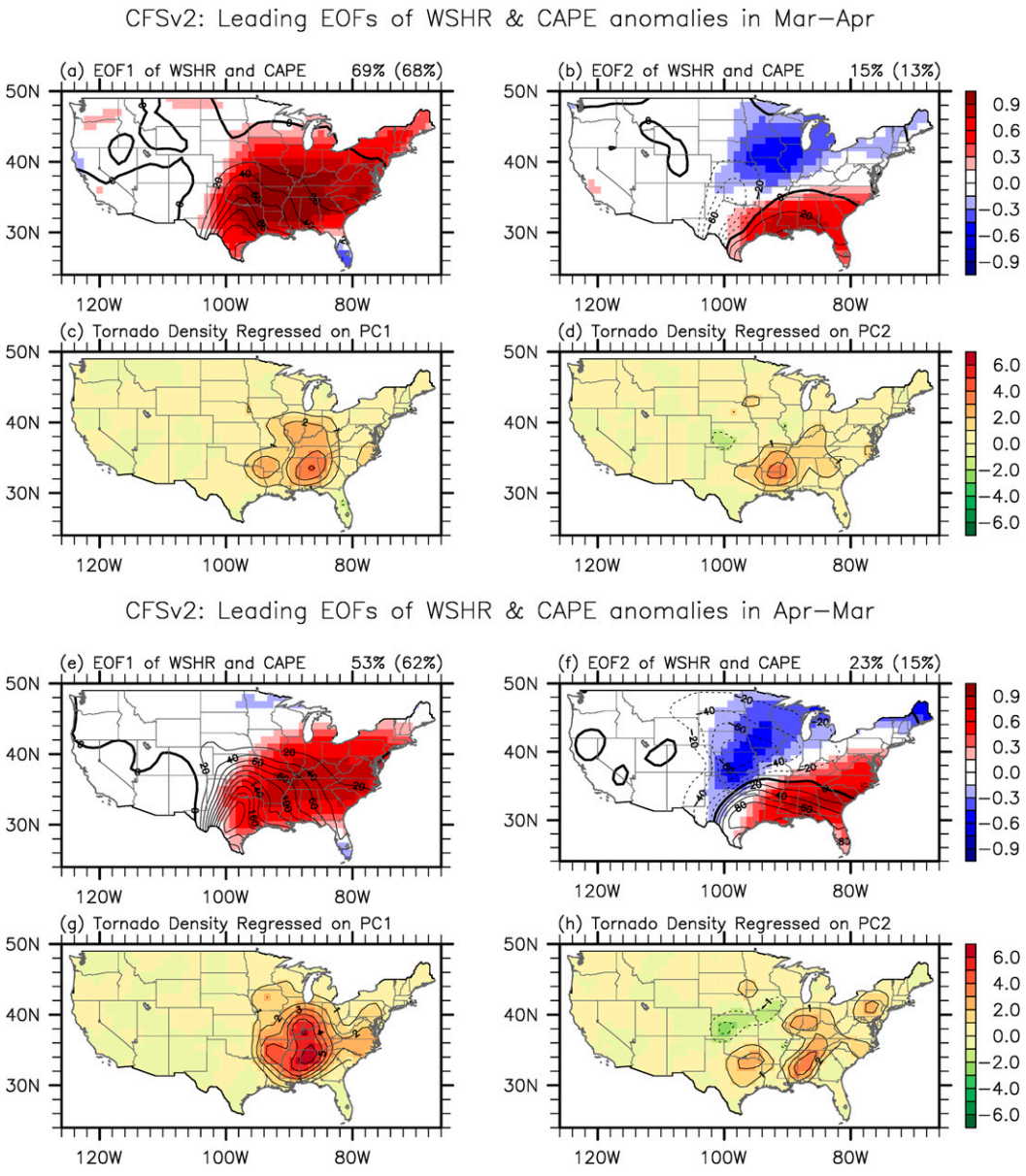

FIG. 3. As in Fig. 2, but using the CFSv2 forecasts. The 20 ensemble members initialized in February of each year during 1982–2018 are merged into a single time series (i.e., 740 samples) to perform the EOF analysis for the target months of MA. Similarly, those initialized in March are used for the target months of AM.

mainly describe dipole-like variability of CAPE and WSHR between the regions northwest and southeast of the Ohio River (Fig. 3b), similar to the two sets of EOF2 derived from ERA-Interim (Figs. 2b,f). They are largely linked to tornado density variability across Arkansas, Louisiana and Mississippi, with weaker variability of opposite sign over Oklahoma in MA (Fig. 3d). A similar tornado density variability pattern is linked to the two sets of EOF2 in AM (Fig. 3h). However, the tornado density variability pattern in AM is more wide spread from the South to the Northeast, and the largest variability is located across Mississippi and Alabama.

The spatial patterns of WSHR and CAPE variability described by the two sets of EOF1 and EOF2 are quite consistent between ERA-Interim and the CFSv2 forecasts. However, the temporal correlations between ERA-Interim and the CFSv2 forecasts for these EOFs are not very strong. For instance, the temporal correlations between ERA-Interim and the CFSv2 forecasts in MA are 0.40 and 0.43 for EOF1 of WSHR and EOF1 of CAPE, respectively. These correlation values are statistically significant at the 99% level based on a Student's *t* test. However, the correlations decrease drastically to 0.09 and 0.07 for EOF2 of WSHR and EOF2 of CAPE, respectively. In AM, the temporal correlations between ERA-Interim and the

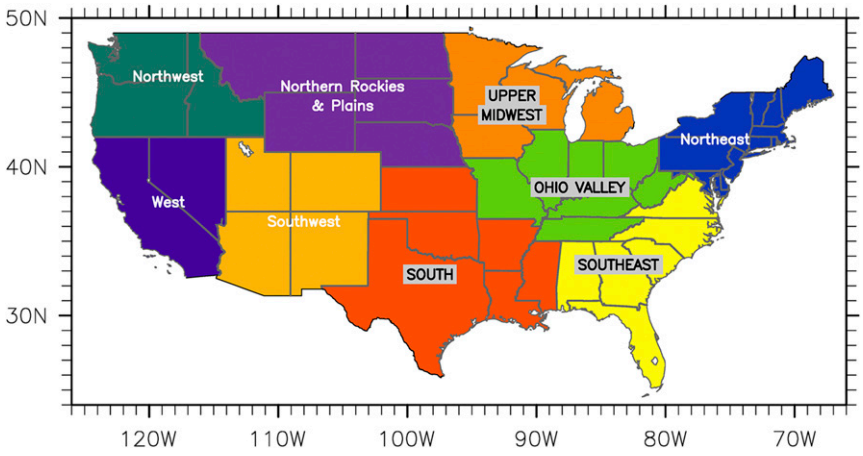

FIG. 4. U.S. climate regions defined by the NCEI. Probabilistic forecast skill of the seasonal forecast model is evaluated for the contiguous United States and for the four U.S climate regions most vulnerable to tornadoes: the Ohio Valley, Southeast, South, and Upper Midwest.

CFSv2 forecasts are 0.51 and 0.10 for EOF1 of WSHR and EOF1 of CAPE, respectively, and 0.20 and 0.13 for EOF2 of WSHR and EOF2 of CAPE, respectively.

Despite the moderate to weak correlations between ERA-Interim and the CFSv2 forecasts, the two sets of EOF1 derived from the CFSv2 forecasts are linked to tornado density variability over multiple U.S. states in the South, Southeast and Ohio valley, while the two sets of EOF2 from the CFSv2 forecasts are largely linked to dipole-like tornado density variability between the southern and central United States. These suggest that the independent sets of EOF1 and EOF2 derived from the CFSv2 forecasts may provide useful predictability of U.S. tornado activity in MA and AM. In the next section, we present the seasonal forecast model, using the independent sets of EOF1 and EOF2 time series from the CFSv2 forecasts as the primary predictors, and test its probabilistic forecast skill for U.S. tornado activity.

4. Probabilistic forecast skill for the contiguous United States and four U.S. climate regions

First, we perform a multiple linear regression analysis with the number of EF2–EF5 tornadoes in MA for the contiguous United States as the dependent variable (i.e., predictand) and the independent sets of EOF1 and EOF2 of WSHR and CAPE for MA derived from the February-initialized CFSv2 forecasts as the independent variables (i.e., predictors). The multiple linear regression analysis is performed for each of the 20 ensemble members. Then, for each of the 20 members, the predicted numbers of MA tornadoes for the period of 1982–2018 are sorted into three categories, namely above-, near-, and below-normal activity. We perform the same analysis by using the number of EF2–EF5 tornadoes in AM for the contiguous United States, and the independent sets of EOF1 and EOF2 of WSHR and CAPE for AM derived from the March-initialized CFSv2 forecasts. The fraction of the ensemble that falls into each of the three categories represents the probability of occurrence. For instance, if 12, 2, and 6 members indicate above-, near-, and below-normal categories, respectively, the probabilistic forecast is 60% of above-normal, 10% of near-normal, and 30% of below-normal activity. The same analyses are repeated and presented for each of the four U.S. climate regions most vulnerable to tornadoes: the Ohio Valley, South, Southeast, and Upper Midwest—see Fig. 4 for the map of the U.S. climate regions, as defined by the National Centers for Environmental Information (NCEI).

For cross validation of the probabilistic reforecast skill, a jackknife resampling technique (e.g., Mosteller and Tukey 1977) is used. Specifically, for each of the 20 ensemble members, the multiple linear regression analysis is repeated by withholding one year of training data, computing the partial regression coefficients using only the remaining 36 years, and then predicting the withheld year. This process is repeated for each of the 37 years, withholding a different year each time, and then the forecast skill for the withheld years is evaluated. The jackknife cross-validation skill can be considered as a lower bound of the prediction skill, while the full-year (including the target year) trained model skill as an upper bound. Therefore, in this section, the skill scores of the full-year trained model for the contiguous United States and the four climate regions are evaluated and compared with the corresponding jackknife cross-validated skill scores. Specifically, we use an ad hoc criteria that the probabilistic reforecast in any category is a useful discriminator between events and non-events (i.e., skillful) only if the ROC score from the full-year trained model is higher than or equal to 0.7, and the ROC score from the jackknife cross-validation test is higher than or equal to 0.6.

Table 2 is a summary of the probabilistic reforecast skill scores for the contiguous United States and the four climate regions. The cross-validated skill scores based on the jackknife tests are also shown in parentheses. For the contiguous United States, the probabilistic reforecast is skillful for the above- and below-normal MA and AM tornado activity. The reforecast is also skillful for the above- and below-normal MA activity in the Ohio Valley and South, the above- and below-normal AM

TABLE 2. RPSS and ROC score values for the three categories (above normal, near normal, and below normal): The RPSS and ROC score values are obtained by using the entire data period (1982–2018) to train the forecast model. The cross-validated RPSS and ROC score values from the jackknife test are also shown in parentheses. For all tests, the number of EF2–EF5 tornadoes is used as the tornado index. The RPSS and ROC score values are in boldface type if RPSS ≥ 0.1 and the cross-validated RPSS > 0.0 or ROC score ≥ 0.7 and the cross-validated ROC score ≥ 0.6.

Initial month	Forecast months	Forecast regions	RPSS (jackknife)	ROCS: Above normal (jackknife)	ROCS: Near normal (jackknife)	ROCS: Below normal (jackknife)
Feb	Mar–Apr	CONUS	**0.25 (0.10)**	**0.81 (0.71)**	0.54 (0.39)	**0.79 (0.75)**
		Ohio Valley	**0.32 (0.15)**	**0.77 (0.65)**	0.62 (0.56)	**0.83 (0.73)**
		South	**0.30 (0.11)**	**0.82 (0.69)**	**0.73 (0.67)**	**0.84 (0.73)**
		Southeast	−0.01 (−0.18)	0.62 (0.48)	0.41 (0.33)	0.65 (0.55)
		Upper Midwest	0.14 (−0.09)	0.72 (0.49)	0.41 (0.40)	0.73 (0.47)
	Mar–May	CONUS	0.13 (−0.02)	0.62 (0.52)	0.56 (0.47)	**0.78 (0.65)**
		Ohio Valley	0.25 (−0.02)	0.71 (0.50)	0.44 (0.36)	**0.83 (0.67)**
		South	0.15 (−0.10)	0.73 (0.55)	0.51 (0.53)	0.67 (0.36)
		Southeast	0.12 (−0.06)	0.68 (0.52)	0.49 (0.45)	**0.71 (0.62)**
		Upper Midwest	0.17 (−0.06)	0.77 (0.48)	0.38 (0.41)	0.69 (0.50)
Mar	Apr–May	CONUS	**0.22 (0.03)**	**0.75 (0.62)**	0.54 (0.38)	**0.82 (0.64)**
		Ohio Valley	**0.27 (0.06)**	0.74 (0.50)	0.61 (0.57)	**0.76 (0.60)**
		South	0.19 (−0.10)	0.81 (0.45)	0.64 (0.48)	0.70 (0.43)
		Southeast	**0.14 (0.04)**	**0.76 (0.68)**	0.28 (0.26)	0.63 (0.57)
		Upper Midwest	**0.31 (0.07)**	**0.88 (0.67)**	0.62 (0.45)	**0.81 (0.62)**
	Apr–Jun	CONUS	0.11 (−0.17)	0.72 (0.43)	0.41 (0.31)	0.64 (0.41)
		Ohio Valley	**0.23 (0.02)**	0.71 (0.49)	0.39 (0.37)	**0.81 (0.64)**
		South	0.16 (−0.17)	0.76 (0.31)	0.36 (0.34)	0.69 (0.40)
		Southeast	0.15 (0.00)	**0.80 (0.62)**	0.47 (0.43)	0.61 (0.51)
		Upper Midwest	0.26 (−0.03)	0.82 (0.52)	0.62 (0.50)	0.80 (0.52)

activity in the Upper Midwest, and the above-normal AM activity in the Southeast. However, the reforecast shows low skill for the above- and below-normal MA activity in the Ohio Valley and South, and the above- and below-normal AM activity in the Southeast and Upper Midwest. In general, there is little to no forecast skill in the February-initialized forecast for MAM, the March-initialized forecast for AMJ or in the near-normal category. In the next subsections, the probabilistic reforecast skills for the contiguous United States and the four climate regions are examined in more detail, focusing on the reforecast skill for the above- and below-normal MA and AM activity.

a. Contiguous United States

Figure 5 summarizes the probabilistic reforecast for the area-averaged tornado activity in the contiguous United States for the above- and below-normal categories and the corresponding reforecast skill metrics (i.e., RPSS; ROC curves and scores). The solid lines are based on the full-year trained model (FYM), whereas the dashed lines are derived from the jackknife model (JKM). The points on the ROC curve indicate the threshold percentage of ensemble members needed to issue a warning for the given category. Starting from the bottom-left corner, the first point indicates the hit rate versus false alarm rate for which all 20 ensemble members are required to be in the tercile (i.e., 100% threshold probability) to issue a warning. The second point away from the bottom-left corner indicates the forecast skill for which 19 of the total 20 (i.e., 95% threshold probability) are required to issue a warning, and so forth. A forecast system that always issues a warning will have hit and false alarm rates equal to one (i.e., perpetual warning or top-right corner), whereas a forecast system that never issues a warning will have hit and false alarm rates equal to zero (i.e., perpetual nonwarning, or bottom-left corner). An ideal forecast system would have relatively high hit rates and low false alarm rate, thus at least some of the points on the ROC curve would lie near the top-left corner of the diagram.

For the target months of MA, the RPSS values for the contiguous United States are 0.25 and 0.10 for FYM and JKM, respectively, which means that the probabilistic reforecast is overall better than the forecast based on climatology (i.e., 33% chance for all three categories). The ROC curves are on the top-left side for both the above- and below-normal categories. Therefore, the ROC scores are higher than 0.5 for both the above-normal (0.81 for FYM and 0.71 for JKM) and below-normal categories (0.79 for FYM and 0.75 for JKM). These ROC scores also meet the requirement (i.e., FYM ≥ 0.7 and JKM ≥ 0.6), thus indicating that the probabilistic reforecasts for these two categories are skillful. For instance, a warning issued at the threshold probability of 50% for the above-normal category (i.e., 10 of 20 ensemble members fall into this category) results in 61% hit rate at the expense of 20% false alarm rate for FYM, and 48% hit rate at the expense of 26% false alarm rate for JKM. Therefore, the hit rate is ~2–3 times the false alarm rate, which indicates that the probabilistic reforecast is useful for this category.

For the target of months of AM, the RPSS values for the contiguous United States are 0.22 and 0.03 for FYM and JKM, respectively. These values are lower than those for MA,

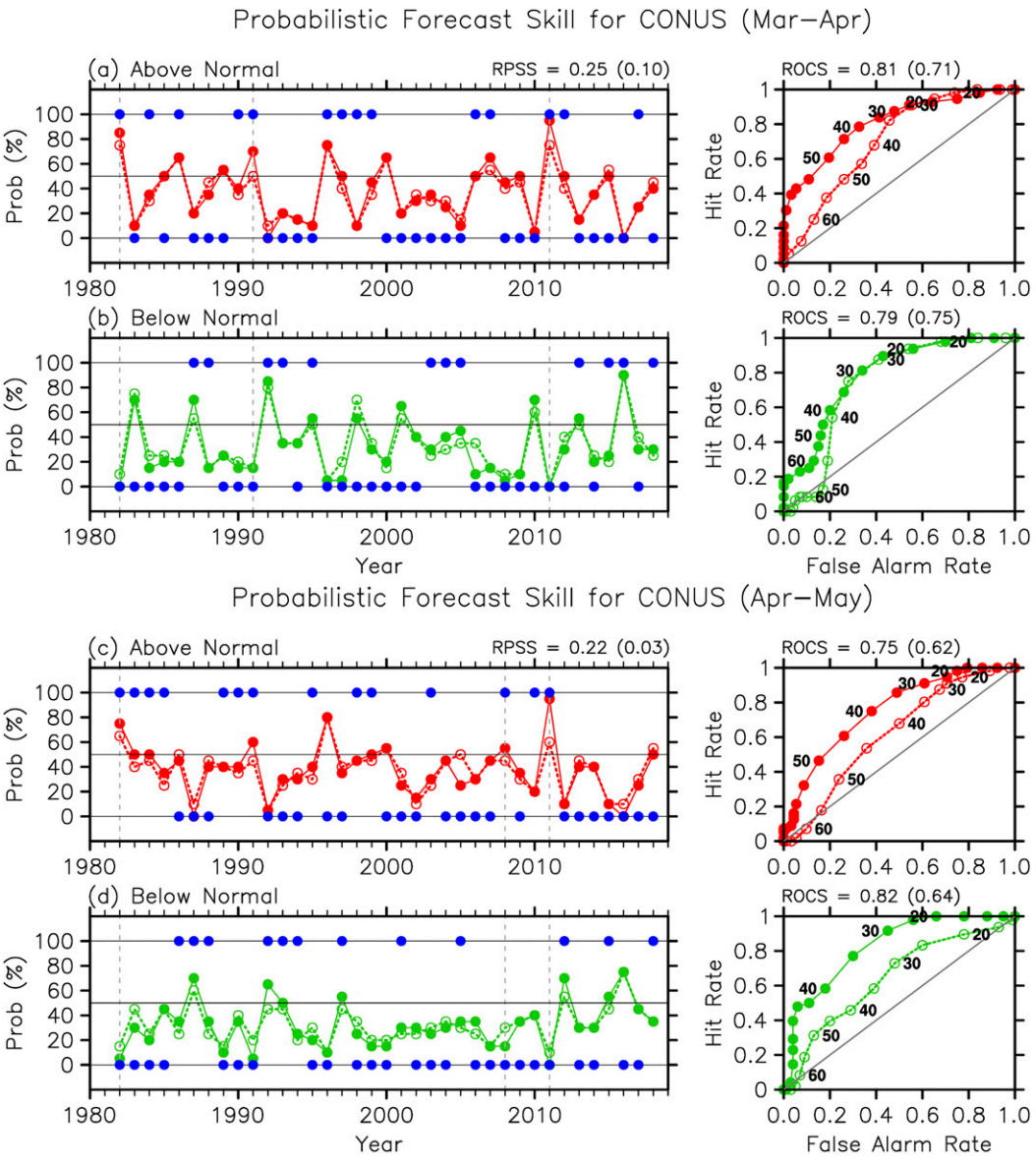

FIG. 5. (left) Probabilistic reforecast of the area-averaged tornado activity in the contiguous United States for the target months of (a),(b) MA and (c),(d) AM, and (right) the corresponding ROC curves for the (top),(bottom middle) above- and (top middle),(bottom) below-normal categories. The solid lines are based on the FYM, whereas the dashed lines are derived from the JKM. The RPSS and ROC scores are indicated for FYM and in parentheses for JKM. The blue dots in the left panels indicate observed events for which the probability is 100% if an incident occurs and 0% if an event does not occur. The gray dashed lines indicate the three most active MA (1982, 1991, and 2011) and the three most active AM (1982, 2008, and 2011) in the contiguous United States. For the ROC curves in the right panels, dots and corresponding numbers along each curve represent different warning thresholds, i.e., the fraction of ensemble members that must fall into that category to issue a warning for that category. The lowest threshold (the top right of each panel) always warns; it hits every actual event but also issues false alarms whenever there is a nonevent. The highest threshold (the bottom left of each panel) never warns; it avoids false alarms but also never hits actual events. A perfect forecast warning system and threshold would lie at the top left of each panel, with a 100% hit rate and no false alarms. An ROC curve for forecasts that are based simply on the observed climatological probability distribution would lie along the diagonal.

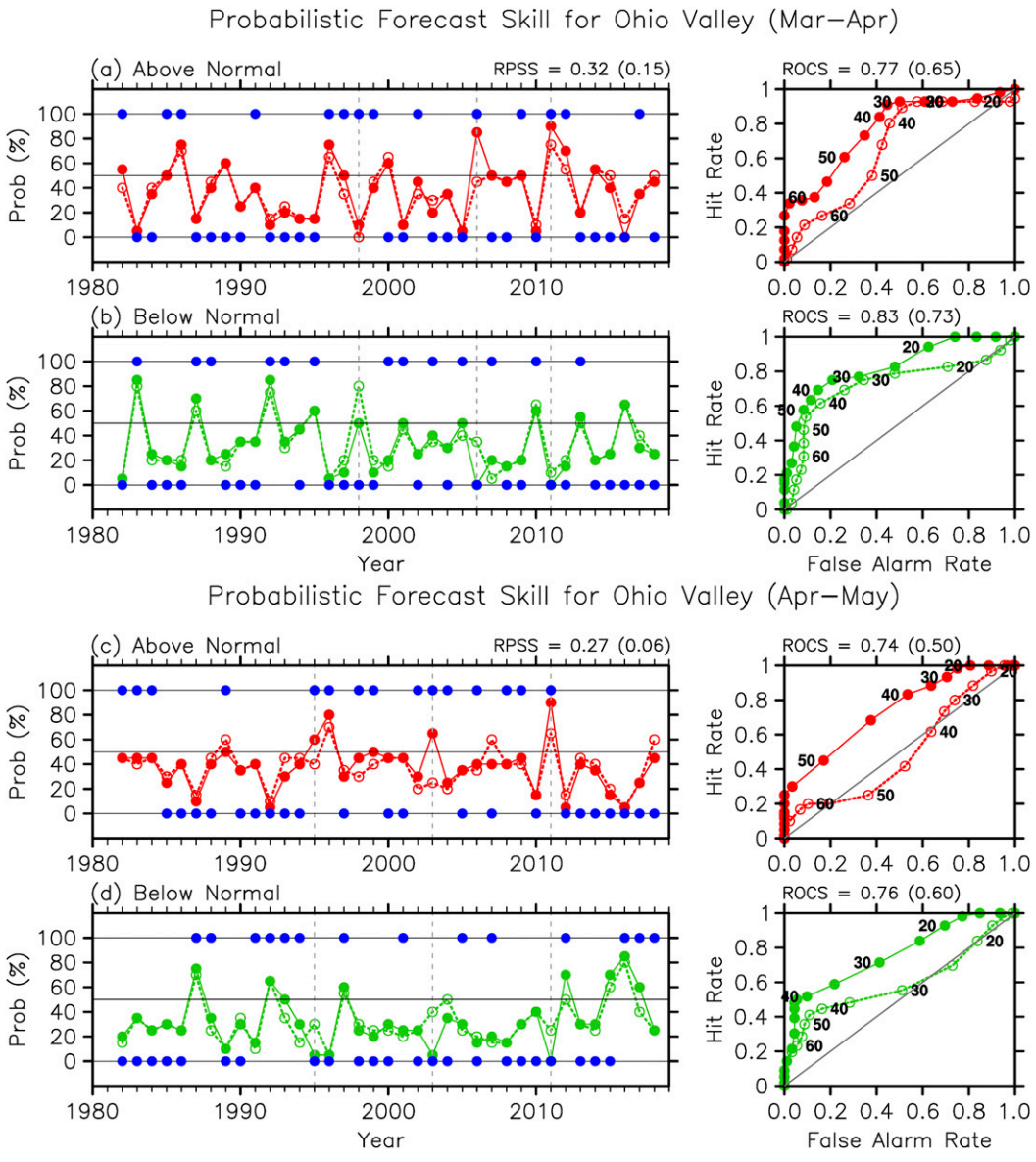

FIG. 6. As in Fig. 5, but for the area-averaged activity in the Ohio Valley. The gray dashed lines indicate the three most active MA (1998, 2006, and 2011) and the three most active AM (1995, 2003, and 2011) in the Ohio Valley.

but still indicate that the probabilistic reforecast is overall better than the forecast based on climatology. The ROC scores are higher than 0.5 for both the above-normal (0.75 for FYM and 0.62 for JKM) and below-normal categories (0.82 for FYM and 0.64 for JKM), and meet the requirement (i.e., FYM $\geq$ 0.7 and JKM $\geq$ 0.6). For instance, a warning issued at the threshold probability of 45% for the above-normal category (i.e., 9 of 20 ensemble members fall into this category) results in 61% hit rate at the expense of 26% false alarm rate for FYM, and 54% hit rate at the expense of 36% false alarm rate for JKM. Therefore, the hit rate is ~1.5–2 times the false alarm rate, indicating that the probabilistic reforecast is useful for this category.

In summary, the probabilistic reforecast for the area-averaged activity in the contiguous United States is skillful for both the February-initialized forecast for MA and the March-initialized forecast for AM, and for both the above-normal and below-normal categories. Consistent with these results, the 2011 super-tornado-outbreak season is successfully predicted by both FYM (95% probability of above-normal activity for both MA and AM) and JKM (75% probability of above-normal activity for MA and 60% for AM), as well as the three other most active U.S. tornado seasons in 1982, 1991, and 2008 at the threshold probability of 50% for MA and 45% for AM (Figs. 5a,c).

b. *Ohio Valley*

Figure 6 shows the probabilistic reforecast for the area-averaged tornado activity in the Ohio Valley for the above- and below-normal categories and the corresponding skill metrics.

For the target months of MA, the RPSS values are 0.32 and 0.15 for FYM and JKM, respectively, indicating that the probabilistic reforecast is more skillful than the reforecast based on climatology. For both the above- and below-normal categories, the ROC curves are in the top-left side, indicating that hit rates are higher than false alarm rates. Consistently, the ROC scores are higher than 0.5 for both the above-normal (0.77 for FYM and 0.65 for JKM) and below-normal categories (0.83 for FYM and 0.73 for JKM), and meet the requirement (i.e., FYM $\geq$ 0.7 and JKM $\geq$ 0.6). For instance, a warning issued at the threshold probability of 40% for the above-normal category (i.e., 8 of 20 ensemble members fall into this category) results in an 84% hit rate at the expense of a 41% false alarm rate for FYM, and an 80% hit rate at the expense of a 46% false alarm rate for JKM. This means that the hit rate is ~2 times the false alarm rate, which makes the probabilistic reforecast useful for this category. For the below-normal category, the ROC curves are clustered toward the bottom-left corner, indicating a very good trade-off between hit rate and false alarm rate.

For the target months of AM, the RPSS values are 0.27 and 0.06 for FYM and JKM, respectively, indicating that the probabilistic reforecast is more skillful than the reforecast based on climatology. For the below-normal category, the ROC curves are in the top-left side, indicating that hit rates are higher than false alarm rates. For the above-normal category, however, the ROC curve for JKM largely follows the diagonal. Therefore, the ROC scores for the above-normal category (0.74 for FYM and 0.50 for JKM) do not meet the requirement. The ROC scores for the below-normal category (0.76 for FYM and 0.60 for JKM) barely meet the minimum requirement. Nevertheless, the ROC curves are clustered toward the lower-left corner for both FYM and JKM, indicating a good trade-off between hit rate and false alarm rate. For instance, a warning issued at the threshold probability of 40% results in a 52% hit rate at the expense of only a 10% false alarm rate for FYM, and a 45% hit rate at the expense of a 16% false alarm rate for JKM. This indicates that the hit rate is ~3–5 times the false alarm rate, which makes the probabilistic reforecast very useful for predicting the below-normal AM activity. This is a good example that shows the value of looking at the shape of the ROC curve and not replying exclusively on the ROC score for decision making.

In summary, the probabilistic reforecast for the Ohio Valley is skillful for predicting the above- and below-normal MA activity and the below-normal AM activity. However, it does not meet the requirement for the above-normal AM activity. Nevertheless, of the five most active regional tornado seasons (1995, 1998, 2003, 2006, and 2011), three seasons, including the 2011 tornado outbreak season, (1995, 2006, and 2011) are successfully predicted by both FYM and JKM at the threshold probability of 40% (Figs. 6a,c).

c. South

For the target months of MA, the RPSS values for the South are 0.31 for FYM and 0.10 for JKM (Fig. 7). These values are slightly smaller than the corresponding RPSS values for the Ohio Valley, but indicate a useful skill relative to the reforecast based on climatology. As in the case of the Ohio Valley, the ROC curves are above the diagonal and the ROC scores meet the requirement for both the above-normal (0.82 for FYM and 0.69 for JKM) and below-normal categories (0.84 for FYM and 0.73 for JKM), indicating that the probabilistic reforecast is useful for those two categories. For instance, a warning issued at the threshold probability of 40% for the above-normal category results in an 88% hit rate at the expense of a 33% false alarm rate for FYM, and a 75% hit rate at the expense of a 41% false alarm rate for JKM. Similarly, a warning issued at the threshold probability of 40% for the below-normal category results in a 71% hit rate at the expense of a 16% false alarm rate for FYM, and a 50% hit rate at the expense of a 20% false alarm rate for FYM. Consistently, three most active MA in the South (i.e., 1982, 1991, and 2011) are successfully predicted at the threshold probability of 40% (Fig. 7a). For the target months of AM, however, the RPSS values are very low, and the ROC scores do not meet the requirement for either the above- or below-normal category. Nevertheless, two of the three most active AM in the South (i.e., 1982 and 2011) are still predicted by both FYM and JKM at the threshold probability of 40% (Fig. 7c).

d. Southeast

For the target months of MA, the RPSS values for the Southeast are −0.01 for FYM and −0.18 for JKM (Fig. 8), indicating that the reforecast is generally not better than a simple reforecast based on climatological probabilities. For both the above- and below-normal categories, the ROC curves for FYM are in the top-left side. However, the corresponding ROC curves for JKM mostly follow the diagonal. Therefore, the ROC scores do not meet the requirement for either the above- or below-normal category. Nevertheless, two of the three most active MA in the Southeast (i.e., 2007 and 2011) are still predicted by both FYM and JKM at the threshold probability of 50% (Fig. 8a).

For the target months of AM, the RPSS values for the Southeast are 0.14 for FYM and 0.03 for JKM (Fig. 8). These values are relatively small, but still indicate a useful skill compared to the reforecast based on climatology. The ROC curves are above the diagonal, and thus the ROC scores are higher than 0.5 for both the above-normal (0.76 for FYM and 0.68 for JKM) and below-normal categories (0.63 for FYM and 0.57 for JKM). However, the ROC scores meet the minimum requirement only for the above-normal category, but not for the below-normal category, indicating that the probabilistic reforecast for AM is skillful only for the above-normal category. For instance, a warning issued at the threshold probability of 45% for the above-normal category results in a 66% hit rate at the expense of a 20% false alarm rate for FYM, and a 63% hit rate at the expense of a 29% false alarm rate for JKM. This indicates that the hit rate is ~2–3 times the false alarm rate, which makes the probabilistic reforecast very useful for the above-normal category. Consistently, three most active AM in the Southeast (i.e., 1989, 2008, and 2011) are successfully predicted by both FYM and JKM at the threshold probability of 45% (Fig. 8c).

e. Upper Midwest

As in the case of the Southeast, for the target months of MA, the RPSS values for the Upper Midwest are very low and the ROC scores do not meet the requirement for either the above- or below-normal category (Figs. 9a,b). Despite the poor skill scores,

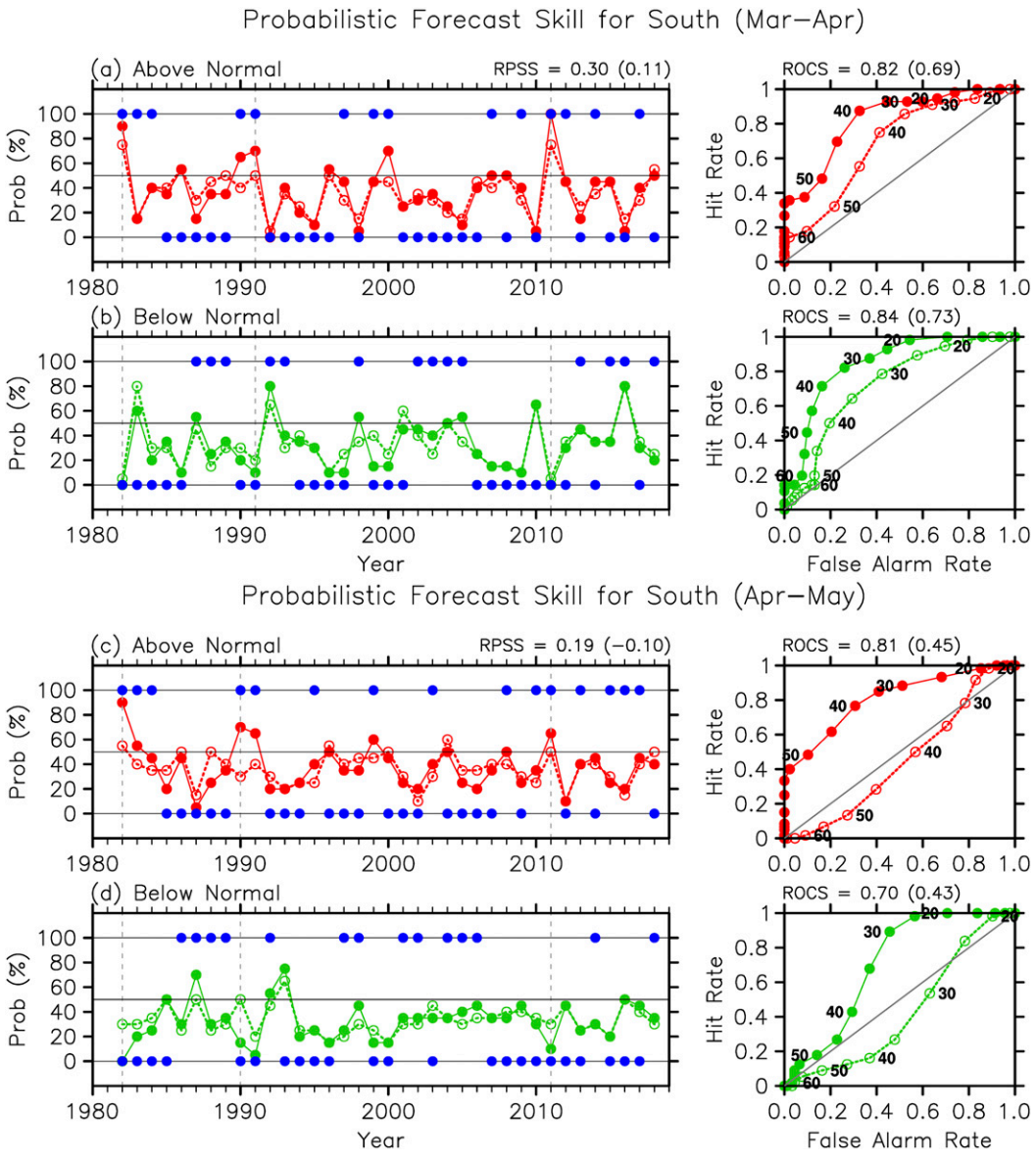

FIG. 7. As in Fig. 5, but for the area-averaged activity in the South. The gray dashed lines indicate the three most active MA (1982, 1991, and 2011) and the three most active AM (1982, 1990, and 2011) in the South.

two of the three most active MA in the Upper Midwest (i.e., 1991 and 2011) are still predicted by both FYM and JKM at the threshold probability of 40% (Fig. 9a). For the target months of AM, on the other hand, the RPSS values for the Southeast are 0.31 for FYM and 0.07 for JKM, indicating a useful skill relative to the reforecast based on climatology. The ROC curves are above the diagonal and the ROC scores meet the requirement for both the above-normal (0.88 for FYM and 0.67 for JKM) and below-normal categories (0.81 for FYM and 0.62 for JKM), indicating that the probabilistic reforecast for AM is skillful for those two categories. For instance, a warning issued at the threshold probability of 40% for the above-normal category results in an 85% hit rate at the expense of a 26% false alarm rate for FYM, and a 65% hit rate at the expense of a 43% false alarm rate for JKM. This indicates that the hit rate is ~1.5–3 times the false alarm rate, which makes the probabilistic reforecast useful for the above-normal category. Consistently, the three most active MA in the Upper Midwest (i.e., 1988, 1991, and 2011) are successfully predicted by both FYM and JKM at the threshold probability of 40% for a warning (Fig. 9c).

5. Probabilistic forecast skill for U.S. regional-scale tornado activity

a. Probabilistic forecast skill metrics

We perform a similar multiple linear regression analysis using the independent sets of EOF1 and EOF2 time series of WSHR and CAPE variability derived from the CFSv2

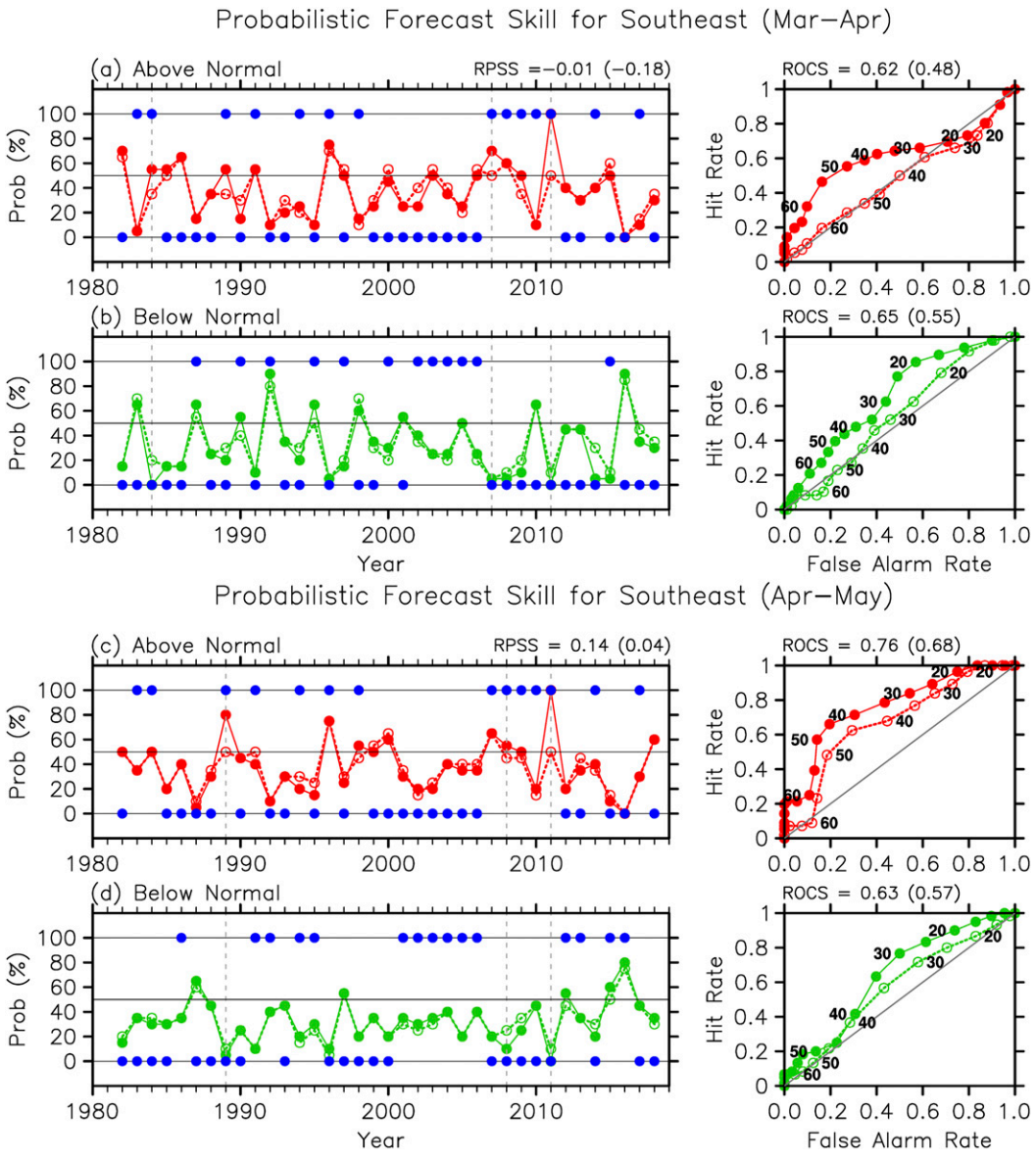

FIG. 8. As in Fig. 5, but for the area-averaged activity in the Southeast. The gray dashed lines indicate the three most active MA (1984, 2007, and 2011) and the three most active AM (1989, 2008, and 2011) in the Southeast.

forecasts as the independent variables (i.e., predictors), but using the tornado density as the dependent variable (i.e., predictand) for each 1° × 1° grid point over the contiguous United States. Figure 10 shows the RPSS values based on FYM. Dark-gray dots indicate where the RPSS values based on JKM are greater than 0, whereas light-gray dots indicate where the RPSS values based on FYM are less than 0.1. The RPSS values based on FYM are mostly above 0.1 in the United States east of 105°W, except in Alabama and Mississippi for the target months of both MA and AM, and the Upper Midwest and Northeast for the target months of MA. However, the positive RPSS values based on JKM for the target months of AM are clustered only in parts of Louisiana, Texas, Kentucky, Tennessee, Iowa, Missouri, and Indiana, and for the target months of MA only in parts of Texas, Oklahoma, Kansas, and several states in the Northeast.

Figure 11 shows the ROC scores based on FYM for the above-, near- and below-normal categories. Dark-gray dots indicate where the ROC scores based on JKM are greater than or equal to 0.6, whereas light-gray dots indicate where the ROC scores based on FYM are less than 0.6. The ROC scores based on FYM are relatively high for both the above- and below-normal categories, but largely less than 0.6 for the near-normal category. For the above-normal category the ROC scores based on JKM for the target months of MA are above or equal to 0.6 (dark-gray dots) over the broad regions around Louisiana and Kentucky and in parts of several states including Texas, Oklahoma, Missouri, Alabama and Georgia. For the

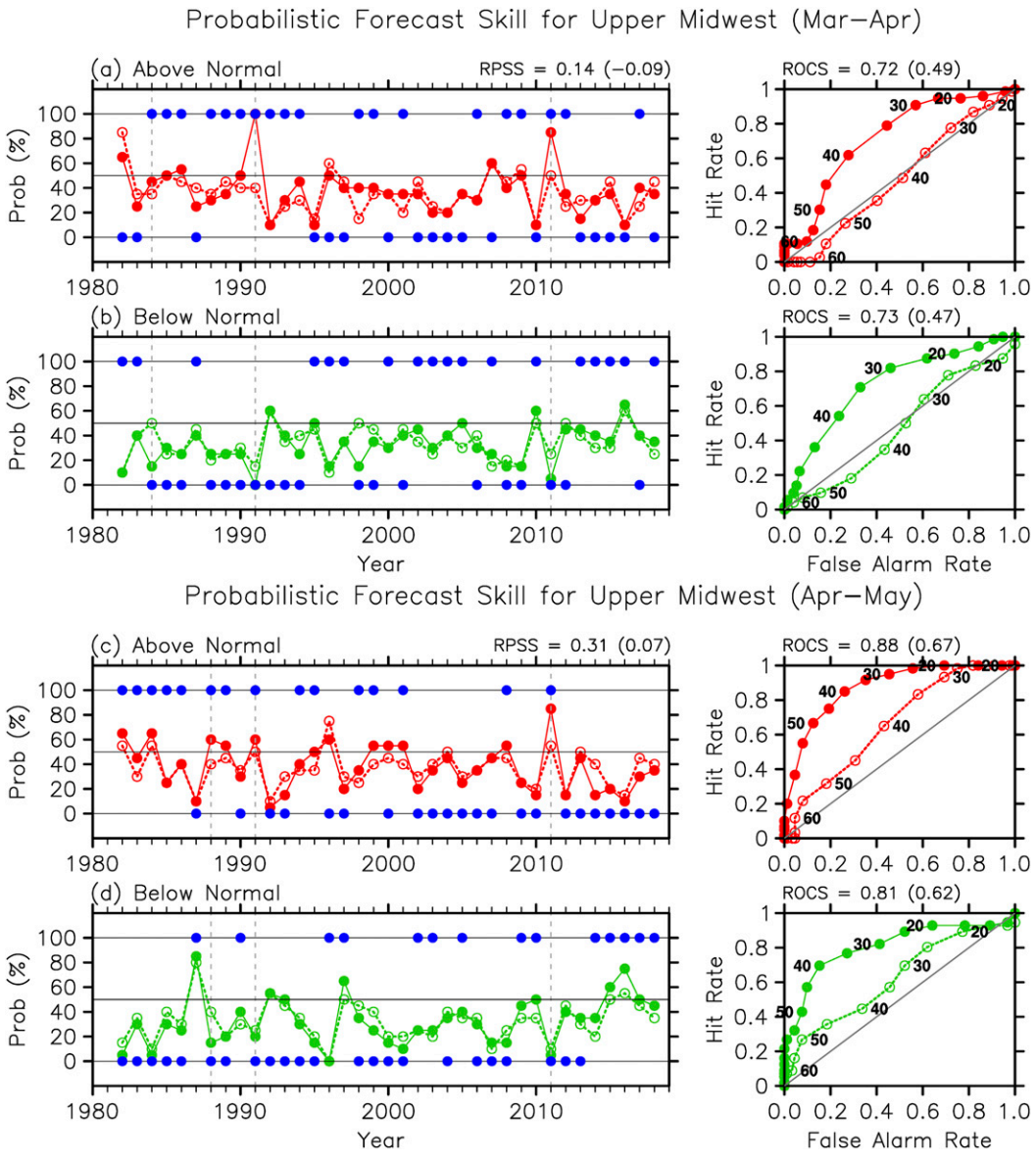

FIG. 9. As in Fig. 5, but for the area-averaged activity in the Upper Midwest. The gray dashed lines indicate the three most active MA (1984, 1991, and 2011) and the three most active AM (1988, 1991, and 2011) in the Upper Midwest.

target months of AM, the regions where the ROC scores based on JKM are above or equal to 0.6 (dark-gray dots) generally shift northward toward Iowa, Michigan, Virginia and North Carolina. Therefore, in the southern United States including Oklahoma, Mississippi, Alabama and Georgia, the ROC scores based on JKM are largely less than 0.6. The spatial distributions of the ROC scores for the below-normal category are quite similar to those for the above-normal category for both MA and MA.

In summary, the probabilistic reforecast for U.S. regional-scale tornado activity is skillful in some regions for the above- and below-normal categories. However, the skill is demonstrated only for either MA or AM in those regions. Additionally, there are many other regions where the reforecast skill is not demonstrated. Therefore, the seasonal outlook for U.S. regional-scale tornado activity based on our method may not yet be ready for an operational use.

b. *Seasonal forecast for the 2011 superoutbreaks*

As illustrated in the previous section, the seasonal forecast model presented in this study cannot be used to accurately forecast year-to-year variability in regional-scale tornado activity, which is greatly affected or driven by synoptic weather patterns, weather regimes and subseasonal processes (e.g., Miller et al. 2020). Instead, this model is designed to forecast large-scale active tornado seasons, such the 2011 superoutbreak

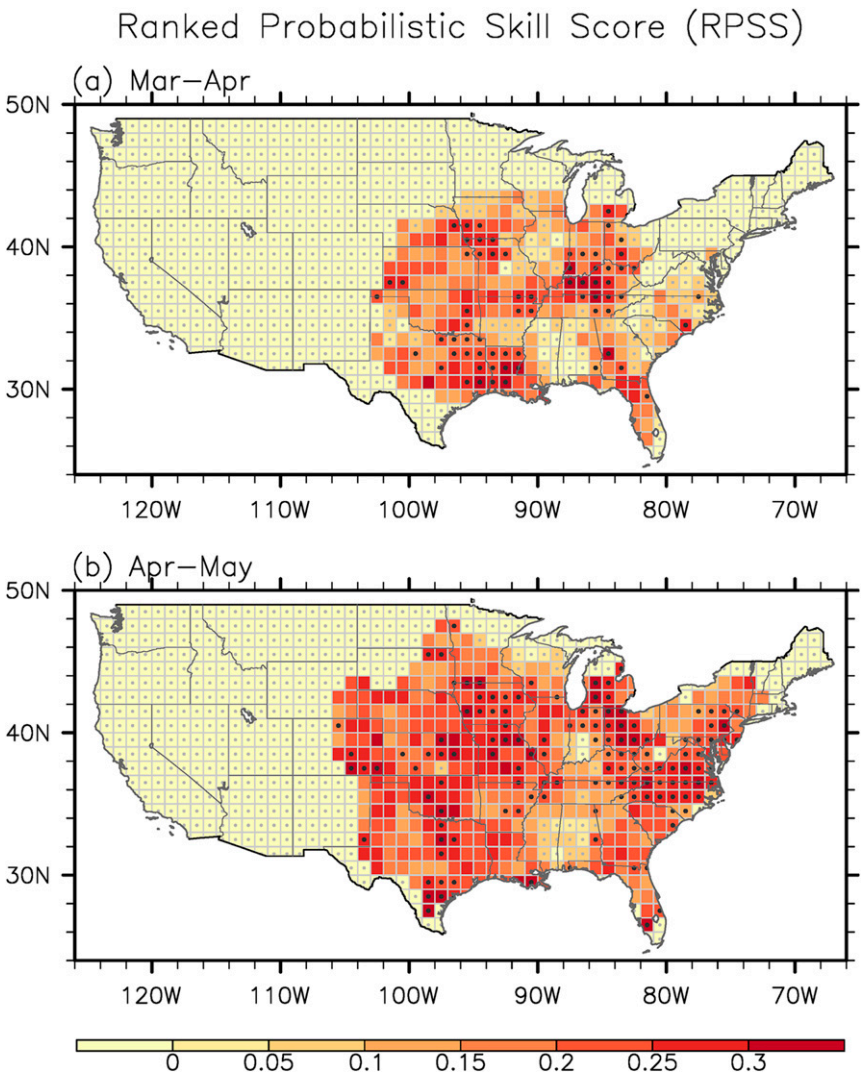

FIG. 10. Local RPSS for the reforecast of U.S. regional-scale tornado activity for the target months of (a) MA and (b) AM from FYM. Dark-gray dots indicate where the RPSS values from FYM are higher than or equal to 0.1 and the scores from JKM are higher than 0.0. Light-gray dots indicate where the RPSS values from FYM are lower than 0.1.

season, that are partly driven by ENSO and other slow-varying ocean and sea ice signals. In that sense, probably the most important test for the seasonal forecast model is to reforecast the 2011 superoutbreak season. As shown in section 4, the 2011 super-tornado-outbreak season is successfully reforecast for the contiguous United States and for each of the four climate regions, for the target months of both MA and AM at the threshold probability of 50%. Here, we further test to explore to what extent the seasonal forecast model can reforecast the regional-scale distribution of U.S. tornado activity of the 2011 season. Figure 12 shows the probabilistic reforecast for U.S. regional-scale tornado activity for the below- (indicated by negative values and green shades) and above-normal (indicated by positive values and red shades) categories, and the validation for the 2011 tornado outbreak season. Gray dots in the left and center panels indicate that the forecast probability is above 50% for either the above- or below-normal category. Note that the threshold tornado density values for the above- and below categories are different in each of the $1° \times 1°$ grid points. For example, a tornado density value of 7 may fall into an above-normal category for a certain grid point, but may fall into a near-normal category in another grid point (Lee et al. 2016).

As shown in Fig. 12 (in the left panels), the probabilistic reforecast based on FYM well captures the above-normal regions of the 2011 super-tornado-outbreak season for the target months of both MA and AM. In particular, the increased tornado activity in the Ohio Valley, South and Southeast is relatively well captured. However, it should be noted that the tornado density data for the 2011 outbreak season are already utilized in the multiple linear regressions to construct FYM. Thus, a more stringent and realistic test is carried out by reconstructing the forecast model with the tornado density data

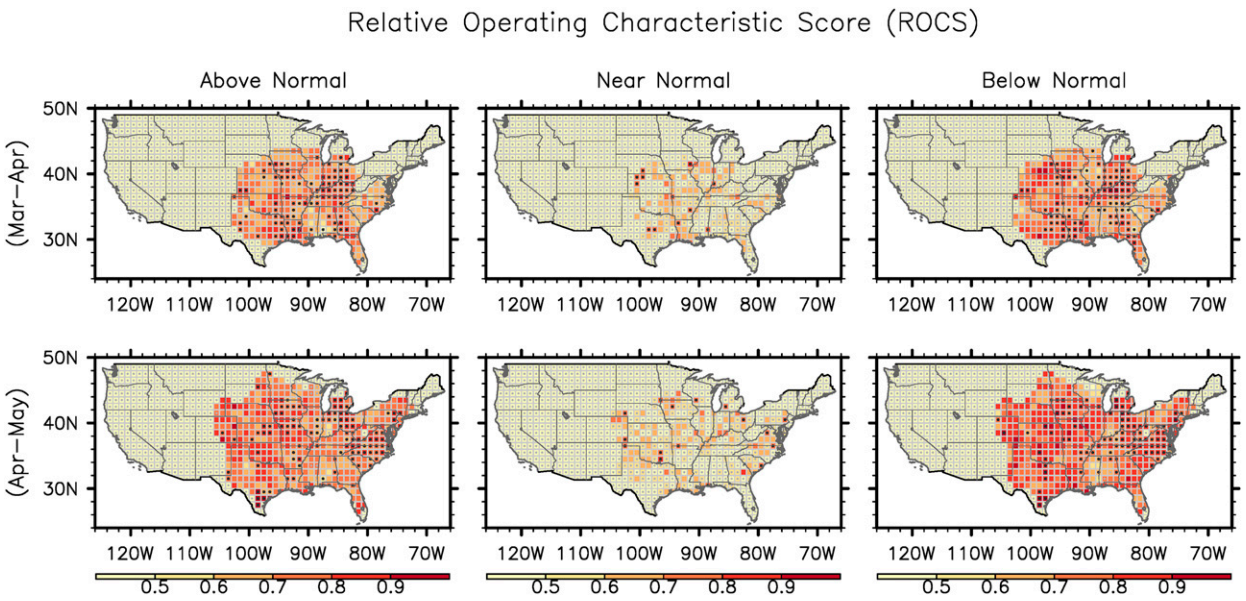

FIG. 11. Local ROC scores for the reforecast of U.S. regional-scale tornado activity for the target months of (top) MA and (bottom) MA for the (left) above-, (center) near-, and (right) below-normal categories. Dark-gray dots indicate where the ROC scores from FYM are higher than equal to 0.7 and the scores from JKM are higher than or equal 0.6. Light-gray dots indicate where the ROC scores from FYM are lower than 0.6.

from 2011 to 2018 excluded. As shown in Fig. 12 (in the center panels), the probabilistic reforecast for MA 2011 based on the revised regression period (1982–2010) still captures the above-normal regions across Oklahoma and Arkansas in the South, Illinois, Indiana, Ohio, Kentucky and Tennessee in the Ohio Valley, and Alabama and Georgia in the Southeast. However, the reforecast for AM 2011 based on the revised regression period only captures a portion of the above-normal regions, mainly over the Southeast, but fails to capture other above-normal regions around Arkansas, Mississippi, Missouri, Wisconsin, Illinois, and most of the Northeast. These results largely confirm that the seasonal outlook for U.S. regional-scale tornado activity based on

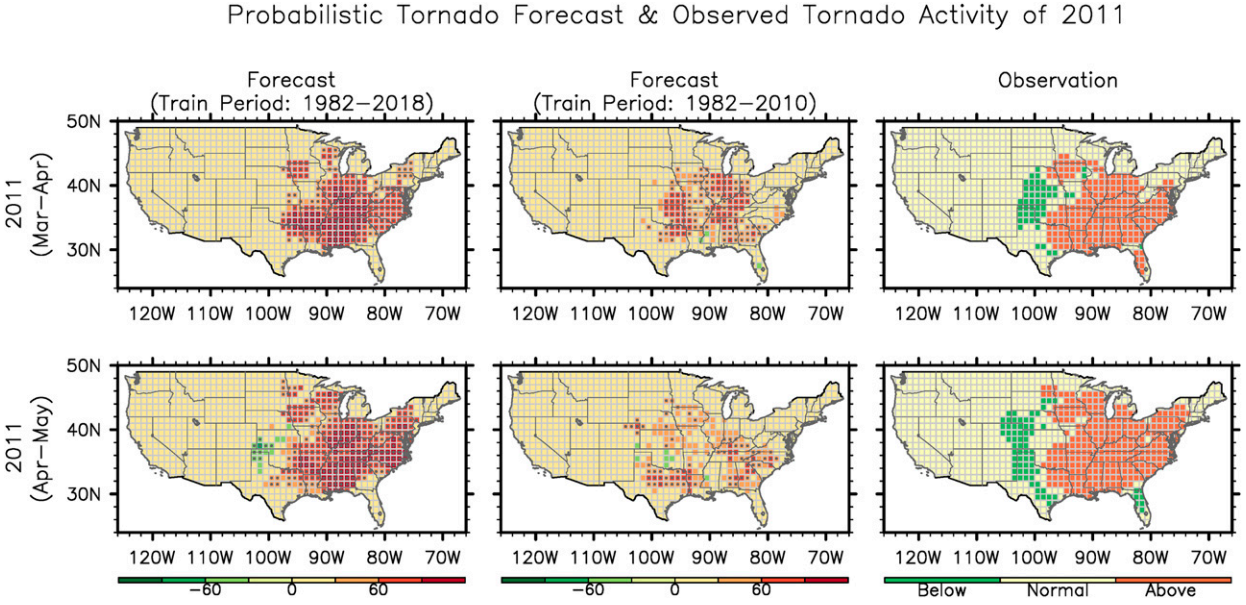

FIG. 12. Probabilistic reforecast for U.S. regional tornado-scale activity for the below-normal (indicated by negative values and green shades) and above-normal (indicated by positive values and red shades) categories for (left) FYM and (center) the revised regression period (1982–2010) and (right) the validation for the 2011 tornado outbreak season for (top) MA and (bottom) AM. Gray dots in the left and center panels indicate that the reforecast probability is above 50%.

our method is not yet suitable for an operational use. Nevertheless, it is very encouraging that the increased regional-scale tornado activity in MA 2011 is fairly well captured over the broad regions in the Ohio Valley, South and Southeast.

6. Discussion

It is interesting to note that the seasonal forecast model shows low to no reforecast skill for the area-averaged MA activity in the Southeast, but shows useful reforecast skill for the area-averaged MA activity in the Ohio Valley and South, and vice versa for the AM activity. To better understand this regional difference in the reforecast skill, it is important to note that the predictability of the current model largely comes from the two sets of EOF1 of WSHR and CAPE variability, which have peak loadings over the Ohio Valley and South (Figs. 2a and 3a). Additionally, the two sets of EOF2 of WSHR and CAPE variability have peak loadings over the Southeast (Figs. 2b and 3b), but their temporal variations are poorly captured by the CFSv2 forecast for MA. For target months of AM, the EOF1 time series of CAPE variability is poorly captured by the CFSv2 forecast, while the EOF2 time series of WSHR variability is better captured comparing to the MA forecast. These help to explain why the seasonal forecast model has useful reforecast skill for the MA activity in the Ohio Valley and South, but low to no skill in the Southeast, and vice versa for the AM activity.

There are other limitations of the experimental seasonal forecast model. The current model is a hybrid model, which uses both a dynamic forecast model and a statistical model. Therefore, it suffers from many issues inherent in the dynamical and statistical forecast models used. For example, the current model heavily relies on historical tornadoes and their links to large-scale tornadic environmental parameters. Since CFSv2 forecasts are only available from 1982 onward, the statistical model is trained for a relatively short period of 37 years (1982–2018). Consequently, if an active tornado season occurs in the future and is not represented by any of the active seasons during the training period, the probabilistic forecast may fail for some regions (e.g., the Southeast for the target months of MA), as demonstrated in the jackknife cross-validation tests. Therefore, there is a need for a global climate reforecast product that goes back to the 1950s, in order to take a full advantage of the historical tornado database.

Another issue is that the probabilistic reforecast skill for the near-normal category is poor. This is commonly observed in many categorical forecasts (e.g., Van den Dool and Toth 1991), and also expected for forecast systems that are partly based on regression (i.e., hybrid dynamical-statistical forecast systems) as in this forecast model. In other words, a regression model, by construction, cannot depict near-normal anomalies very well. Another way to interpret the poor reforecast skill for the near-normal category is that the absence of climate signals in the forecast does not necessarily imply near-normal U.S. tornado activity. In other words, active or inactive tornado seasons may occur due to unpredictable weather events, even when the seasonal forecast model predicts near-normal conditions for the tornadic environmental parameters.

Another potential problem originates from the CFSv2 forecasts. As discussed in section 3, the spatial patterns of WSHR and CAPE variability described by the two sets of EOF1 and EOF2 are very similar between the ERA-Interim reanalysis and the CFSv2 forecasts. However, the temporal correlations of the two sets of EOF2 between ERA-Interim and the CFSv2 forecasts are weak, indicating that the skill in predicting large-scale seasonal WSHR and CAPE variability may currently limit the predictions of U.S. tornado activity. Future work should test whether this issue applies to other dynamic seasonal forecast systems participating in the North American Multi-Model Ensemble (Infanti and Kirtman 2014; Kirtman et al. 2014; Becker and van den Dool 2016), and if it can be improved by using a multimodel ensemble approach.

Earlier studies have shown the ENSO-induced extratropical teleconnection patterns over the United States can be greatly modulated by the Madden–Julian oscillation (MJO) at the subseasonal time scale (i.e., 14–30 days). In particular, the extratropical response is enhanced when the MJO- and ENSO-induced tropical convections are in-phase, and weakened when they are out-of-phase (e.g., Roundy et al. 2010; Moon et al. 2011; Riddle et al. 2013; Johnson et al. 2014; Arcodia et al. 2020). Therefore, it may be useful to explore subseasonal predictability of U.S. tornado activity in MAM using various potential predictors such as tropical atmospheric convective activity associated with the MJO, global atmospheric angular momentum oscillations, and regional weather regimes (e.g., Thompson and Roundy 2013; Barrett and Gensini 2013; Gensini and Marinaro 2016; Tippett 2018; Baggett et al. 2018; Moore and McGuire 2020; Gensini et al. 2019, Kim et al. 2020; Miller et al. 2020). For example, a recent study (Kim et al. 2020) showed a promising result that subseasonal U.S. tornado activity in May-June-July is strongly tied to certain phases of the MJO and associated convective activity across the northeast Pacific and Central America.

7. Summary and conclusions

This study describes an experimental model for Seasonal Probabilistic Outlook for Tornadoes (SPOTter) in the contiguous United States for the target season of MAM. We test the probabilistic forecast skill by using various statistical measures including the RPSS, and ROC curve and score. The independent sets of EOF1 and EOF2 of WSHR and CAPE variability over the contiguous United States are obtained from the CFSv2 forecasts and used as the primary predictors, with the premise that the modulating impacts of ENSO and other slowly varying ocean and sea ice processes are integrated into these predictors. The initial forecast is carried out using the February-initialized CFSv2 forecasts for the target months of MA, and then updated using the March-initialized CFSv2 forecasts for the target months of AM. A series of comprehensive cross-validation reforecast skill tests for the period of 1982–2018 shows that the probabilistic reforecast is skillful in predicting the area-averaged tornado activity over the contiguous United States for the above- and below-normal categories for the target months of both MA and AM. Consistently, the 2011 super-tornado-outbreak season as well as the three other

most active U.S. tornado seasons in 1982, 1991, and 2008 are successfully reforecast. Therefore, the probabilistic forecast model presented in this study may be suitable for an operational use in predicting future active and inactive U.S. tornado seasons.

Additional skill tests applied to the four U.S. climate regions show that the probabilistic reforecast successfully captures the 2011 outbreak season in all four climate regions, and is skillful for the area-averaged tornado activity in the Ohio Valley and South for the target months of MA, and in the Southeast and Upper Midwest for the target months of AM, particularly for the above-normal category. However, the probabilistic reforecast skill is poor for predicting the area-averaged tornado activity in the Ohio Valley and the South for the target months of AM, and in the Southeast and Upper Midwest for the target months of MA. Consistent with these results, the probabilistic reforecast skill for U.S. regional-scale tornado activity is demonstrated only in certain regions for the target months of either MA or AM. Therefore, although the reforecast using the model trained for 1982–2010 fairly well captures the increased regional-scale tornado activity in MA 2011 over the Ohio Valley, South and Southeast, the seasonal outlook for U.S. regional-scale tornado activity based on our method (i.e., CFSv2-based hybrid dynamic-statistical forecast) may not yet be ready for an operational use.

Acknowledgments. We thank two anonymous reviewers and Paul Roundy for their insightful comments and suggestions, which led to a significant improvement of the paper. We also acknowledge Altug Aksoy, Israel Jirak, and Adam Clark for helpful comments and suggestions and John Allen, Gerry Bell, Victor Gensini, and Hui Wang for useful discussions during NOAA Climate Prediction Center's seasonal severe weather outlook teleconferences. This work was supported by NOAA Oceanic and Atmospheric Research Grant (03-02-06-011), NOAA Climate Program Office MAPP Grant (NA19OAR4310282), and the NOAA Atlantic Oceanographic and Meteorological Laboratory.

REFERENCES

Allen, J. T., M. K. Tippett, and A. H. Sobel, 2015: Influence of the El Niño/Southern Oscillation on tornado and hail frequency in the United States. *Nat. Geosci.*, **8**, 278–283, https://doi.org/10.1038/ngeo2385.

——, 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.

Arcodia, M. C., B. P. Kirtman, and L. S. P. Siqueira, 2020: How MJO teleconnections and ENSO interference impacts U.S. precipitation. *J. Climate*, **33**, 4621–4640, https://doi.org/10.1175/JCLI-D-19-0448.1.

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.

Becker, E., and H. van den Dool, 2016: Probabilistic seasonal forecasts in the North American Multimodel Ensemble: A baseline skill assessment. *J. Climate*, **29**, 3015–3026, https://doi.org/10.1175/JCLI-D-14-00862.1.

Brooks, H. E., 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.

Childs, S. J., R. S. Schumacher, and J. T. Allen, 2018: Cold-season tornadoes: Climatological and meteorological insights. *Wea. Forecasting*, **33**, 671–691, https://doi.org/10.1175/WAF-D-17-0120.1.

Chu, J. E., A. Timmermann, and J. Y. Lee, 2019: North American April tornado occurrences linked to global sea surface temperature anomalies. *Sci. Adv.*, **5**, eaaw9950, https://doi.org/10.1126/sciadv.aaw9950.

Cook, A. R., and J. T. Schaefer, 2008: The relation of El Niño–Southern Oscillation (ENSO) to winter tornado outbreaks. *Mon. Wea. Rev.*, **136**, 3121–3137, https://doi.org/10.1175/2007MWR2171.1.

——, L. M. Leslie, D. B. Parsons, and J. T. Schaefer, 2017: The impact of the El Niño–Southern Oscillation (ENSO) on winter and early spring U.S. tornado outbreaks. *J. Appl. Meteor. Climatol.*, **56**, 2455–2478, https://doi.org/10.1175/JAMC-D-16-0249.1.

Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, https://doi.org/10.1002/qj.828.

Elsner, J. B., and H. M. Widen, 2014: Predicting spring tornado activity in the central Great Plains by 1 March. *Mon. Wea. Rev.*, **142**, 259–267, https://doi.org/10.1175/MWR-D-13-00014.1.

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.

——, 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.

Hamill, T. M., 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.

Harvey, L. O., Jr., K. R. Hammond, C. M. Lusk, and E. F. Mross, 1992: The application of signal detection theory to weather forecasting behavior. *Mon. Wea. Rev.*, **120**, 863–883, https://doi.org/10.1175/1520-0493(1992)120<0863:TAOSDT>2.0.CO;2.

Infanti, J. M., and B. P. Kirtman, 2014: Southeastern U.S. rainfall prediction in the North American Multimodel Ensemble. *J. Hydrometeor.*, **15**, 529–550, https://doi.org/10.1175/JHM-D-13-072.1.

Johnson, N. C., D. C. Collins, S. B. Feldstein, M. L. L'Heureux, and E. E. Riddle, 2014: Skillful wintertime North American temperature forecasts out to 4 weeks based on the state of ENSO and the MJO. *Wea. Forecasting*, **29**, 23–38, https://doi.org/10.1175/WAF-D-13-00102.1.

Jung, E., and B. P. Kirtman, 2016: Can we predict seasonal changes in high impact weather in the United States? *Environ. Res. Lett.*, **11**, 074018, https://doi.org/10.1088/1748-9326/11/7/074018.

Kharin, V. V., and F. W. Zwiers, 2003: On the ROC score of probability forecasts. *J. Climate*, **16**, 4145–4150, https://doi.org/10.1175/1520-0442(2003)016<4145:OTRSOP>2.0.CO;2.

Kim, D., S.-K. Lee, and H. Lopez, 2020: Madden–Julian oscillation–induced suppression of northeast Pacific convection increases U.S. tornadogenesis. *J. Climate*, **33**, 4927–4939, https://doi.org/10.1175/JCLI-D-19-0992.1.

Kirtman, B. P., and Coauthors, 2014: The North American Multimodel Ensemble: Phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. *Bull. Amer. Meteor. Soc.*, **95**, 585–601, https://doi.org/10.1175/BAMS-D-12-00050.1.

Lee, S.-K., R. Atlas, D. Enfield, C. Wang, and H. Liu, 2013: Is there an optimal ENSO pattern that enhances large-scale atmospheric processes conducive to tornado outbreaks in the United States? *J. Climate*, **26**, 1626–1642, https://doi.org/10.1175/JCLI-D-12-00128.1.

——, A. T. Wittenberg, D. B. Enfield, S. J. Weaver, C. Wang, and R. Atlas, 2016: U.S. regional tornado outbreaks and their links to spring ENSO phases and North Atlantic SST variability. *Environ. Res. Lett.*, **11**, 044008, https://doi.org/10.1088/1748-9326/11/4/044008.

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.

Lopez, H., and B. P. Kirtman, 2014: WWBs, ENSO predictability, the spring barrier and extreme events. *J. Geophys. Res. Atmos.*, **119**, 10 114–10 138, https://doi.org/10.1002/2014JD021908.

Marsh, P. T., H. E. Brooks, and D. J. Karoly, 2007: Assessment of the severe weather environment in North America simulated by a global climate model. *Atmos. Sci. Lett.*, **8**, 100–106, https://doi.org/10.1002/asl.159.

Marzban, C., and J. Schaefer, 2001: The correlation between U.S. tornados and Pacific sea surface temperatures. *Mon. Wea. Rev.*, **129**, 884–895, https://doi.org/10.1175/1520-0493(2001)129<0884:TCBUST>2.0.CO;2.

Mason, S. J., and N. E. Graham, 1999: Conditional probabilities, relative operating characteristics, and relative operating levels. *Wea. Forecasting*, **14**, 713–725, https://doi.org/10.1175/1520-0434(1999)014<0713:CPROCA>2.0.CO;2.

Miller, D. E., 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.

Molina, M. J., and J. T. Allen, 2019: On the moisture origins of tornadic thunderstorms. *J. Climate*, **32**, 4321–4346, https://doi.org/10.1175/JCLI-D-18-0784.1.

——, R. P. Timmer, and J. T. Allen, 2016: Importance of the Gulf of Mexico as a climate driver for U.S. severe thunderstorm activity. *Geophys. Res. Lett.*, **43**, 12 295–12 304, https://doi.org/10.1002/2016GL071603.

——, J. T. Allen, and V. A. Gensini, 2018: The Gulf of Mexico and ENSO influence on subseasonal and seasonal CONUS winter tornado variability. *J. Appl. Meteor. Climatol.*, **57**, 2439–2463, https://doi.org/10.1175/JAMC-D-18-0046.1.

Moon, J. Y., B. Wang, and K. J. Ha, 2011: ENSO regulation of MJO teleconnection. *Climate Dyn.*, **37**, 1133–1149, https://doi.org/10.1007/s00382-010-0902-3.

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.

Mosteller, F., and J. W. Tukey, 1977: *Data Analysis and Regression: A Second Course in Statistics*. Addison-Wesley, 586 pp.

Muñoz, E., and D. E. Enfield, 2011: The boreal spring variability of the Intra-Americas low-level jet and its relation with precipitation and tornadoes in the eastern United States. *Climate Dyn.*, **36**, 247–259, https://doi.org/10.1007/s00382-009-0688-3.

Riddle, E. E., M. B. Stoner, N. C. Johnson, M. L. L'Heureux, D. C. Collins, and S. B. Feldstein, 2013: The impact of the MJO on clusters of wintertime circulation anomalies over the North American region. *Climate Dyn.*, **40**, 1749–1766, https://doi.org/10.1007/s00382-012-1493-y.

Roundy, P. E., K. MacRitchie, J. Asuma, and T. Melino, 2010: Modulation of the global atmospheric circulation by combined activity in the Madden–Julian oscillation and the El Niño–Southern Oscillation during boreal winter. *J. Climate*, **23**, 4045–4059, https://doi.org/10.1175/2010JCLI3446.1.

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.

Saide, P. E., and Coauthors, 2015: Central American biomass burning smoke can increase tornado severity in the U.S. *Geophys. Res. Lett.*, **42**, 956–965, https://doi.org/10.1002/2014GL062826.

Swets, J. A., 1973: The relative operating characteristic in psychology. *Science*, **182**, 990–1000, https://doi.org/10.1126/science.182.4116.990.

Thompson, D. B., and P. E. Roundy, 2013: The relationship between the Madden–Julian Oscillation and U.S. violent tornado outbreaks in the spring. *Mon. Wea. Rev.*, **141**, 2087–2095, https://doi.org/10.1175/MWR-D-12-00173.1.

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.

——, A. H. Sobel, and S. J. Camargo, 2012: Association of U.S. tornado occurrence with monthly environmental parameters. *Geophys. Res. Lett.*, **39**, L02801, https://doi.org/10.1029/2011GL050368.

Trapp, R. J., and K. A. Hoogewind, 2018: Exploring a possible connection between U.S. tornado activity and Arctic sea ice. *npj Climate Atmos. Sci.*, **1**, 14, https://doi.org/10.1038/s41612-018-0025-9.

Van den Dool, H. M., and Z. Toth, 1991: Why do forecasts for "near normal" often fail? *Wea. Forecasting*, **6**, 76–85, https://doi.org/10.1175/1520-0434(1991)006<0076:WDFFNO>2.0.CO;2.

Verbout, S. M., H. E. Brooks, L. M. Leslie, and D. M. Schultz, 2006: Evolution of the U.S. tornado database: 1954–2003. *Wea. Forecasting*, **21**, 86–93, https://doi.org/10.1175/WAF910.1.

Weaver, S. J., S. Baxter, and A. Kumar, 2012: Climatic role of North American low-level jets on U.S. regional tornado activity. *J. Climate*, **25**, 6666–6683, https://doi.org/10.1175/JCLI-D-11-00568.1.