Probabilistic Skill of Subseasonal Surface Temperature Forecasts over North America

N. VIGAUD

International Research Institute for Climate and Society, Earth Institute at Columbia University, Palisades, New York

M. K. TIPPETT

Department of Applied Physics and Applied Mathematics, Columbia University, New York, New York

J. YUAN, A. W. ROBERTSON, AND N. ACHARYA

International Research Institute for Climate and Society, Earth Institute at Columbia University, Palisades, New York

(Manuscript received 11 June 2019, in final form 6 August 2019)

ABSTRACT

The skill of surface temperature forecasts up to 4 weeks ahead is examined for weekly tercile category probabilities constructed using extended logistic regression (ELR) applied to three ensemble prediction systems (EPSs) from the Subseasonal-to-Seasonal (S2S) project (ECMWF, NCEP, and CMA), which are verified over the common period 1999–2010 and averaged with equal weighting to form a multimodel ensemble (MME). Over North America, the resulting forecasts are characterized by good reliability and varying degrees of sharpness. Skill decreases after two weeks and from winter to summer. Multimodel ensembling damps negative skill that is present in individual forecast systems, but overall, does not lead to substantial skill improvement compared to the best (ECMWF) model. Spatial pattern correction is implemented by projecting the ensemble mean temperatures neighboring each grid point onto Laplacian eigenfunctions, and then using those amplitudes as new predictors in the ELR. Forecasts and skill improve beyond week 2, when the ELR model is trained on spatially averaged temperature (i.e., the amplitude of the first Laplacian eigenfunction) rather than the gridpoint ensemble mean, but not at shorter leads. Forecasts are degraded when adding more Laplacian eigenfunctions that encode additional spatial details as predictors, likely due to the short reforecast sample size. Forecast skill variations with ENSO are limited, but MJO relationships are more pronounced, with the highest skill during MJO phase 3 up to week 3, coinciding with enhanced forecast probabilities of above-normal temperatures in winter.

1. Introduction

In comparison to seasonal hindcasts (reforecasts), submonthly hindcasts are often characterized by shorter length and fewer ensemble members, so a straightforward computing of probabilities by counting of ensemble members exceeding a chosen threshold leads to large errors. In the cases of the 4-member NCEP and CMA reforecasts archived in the S2S database and used in this study, for instance, the reforecast probability obtained by counting can only take the values of 0%, 25%, 50%, 75%, and 100%, which is very crude. Distributional regression is, by contrast, well suited to probability forecasting, and regression models are more skillful than straight counting for small ensemble sizes in the seasonal forecasting context (Tippett et al. 2007). Model output statistics (MOS) has been shown to improve probabilistic weather forecasts (Hamill et al. 2004), but fewer analyses have been yet done at subseasonal time scales. Submonthly forecasts based on extended logistic regression (ELR; Wilks 2009) have recently provided probabilistic precipitation forecast skill estimates on S2S time scales over different parts of the globe including North America (Vigaud et al. 2017a,b, 2018), but such approaches have yet to be applied to surface temperatures. In the ELR methodology proposed in Vigaud et al. (2017a), calibration is done at the gridpoint level (i.e., a separate regression model is constructed for every location without using information from neighboring grid points). Since gridpoint regressions are prone to sampling uncertainties that can translate into spatially noisy forecasts, there might be potential for improvements by including spatial information. This study thus aims at providing probabilistic skill estimates for North

Corresponding author: N. Vigaud, nicolas.vigaud@gmail.com

DOI: 10.1175/WAF-D-19-0117.1

TABLE 1. Attributes from ECMWF, NCEP, and CMA forecasts archived in the S2S database at ECMWF.

Attributes	ECMWF	NCEP	CMA
Time range	d0–46	d0–44	d0–60
Resolution	Tco639/319 L91	T126L64	T106L40
Ensemble size	51	16	4
Frequency	2 per week	Daily	Daily
Reforecasts (RFC)	On the fly	Fix	Fix
RFC length	Past 20yrs	1999–2010	1994–2014
RFC frequency	2 per week	Daily	Daily
RFC size	11	4	4

American surface temperature terciles from submonthly reforecasts and examining if these can be improved by multimodel ensembling and spatial pattern correction.

Multiple linear regressions like principal component regressions (PCR; Mo and Straus 2002) or canonical correlation analysis (CCA; Barnston and Ropelewski 1992), are well suited for MOS and the treatment of systematic errors in the positions and amplitudes of patterns in dynamical model seasonal predictions (Ward and Navarra 1997; Rukhovets et al. 1998; Smith and Livezey 1999; Feddersen et al. 1999; Tippett et al. 2003; Barnston and Tippett 2017). However, converting linear regression forecasts into probability forecasts usually requires a Gaussian assumption that may be less appropriate at subseasonal time scales. The pattern-based MOS method often used empirical orthogonal functions (EOFs), which again depend on the data used to develop them, and hence vary by model. By contrast, Laplacian eigenfunction decomposition, which has been recently applied to climate analysis (Saito 2008; DelSole and Tippett 2015), makes no assumption on the distribution of the data and is well suited for multimodel studies because Laplacian eigenfunctions are uniformly defined across models (DelSole and Tippett 2015). The Laplacian eigenfunctions are ordered by length scale from longest to shortest, and thus represent an attractive approach for filtering out small-scale variability and summarizing spatial information. The skill of weekly temperature tercile probability forecasts is first examined by applying ELR at each grid point to each individual models' forecasts separately. The probabilities of the individual models are averaged with equal weighting to form a multimodel ensemble (MME) forecast. Spatial pattern correction is applied through the decomposition of ensemble mean temperature neighboring each grid point using locally defined Laplacian eigenfunctions.

The methods and data are presented in section 2. The skill of weekly forecasts initialized during DJF (winter) and JJA (summer) are examined over North America in section 3. Improvements to skill through spatial pattern correction based on Laplacian eigenfunctions are then discussed with skill relationships to ENSO conditions and MJO phases. Summary and conclusions are gathered in section 4.

2. Data and methods

a. Observation and model datasets

Week-1 through week-4 [i.e., from $(d + 1; d + 7)$ to $(d + 22; d + 28)$ targets for a forecast on day d] daily surface temperatures from the European Centre for Medium-Range Weather Forecasts (ECMWF), National Centers for Environmental Prediction (NCEP), and the China Meteorological Administration (CMA) were all acquired from the S2S database (Vitart et al. 2017) as in Vigaud et al. (2017a), which the following data description parallels in this paragraph. As shown in Table 1, these EPSs have differing resolutions, ensemble size, and reforecasts lengths. The common factor in the S2S database is that they are all archived on the same 1.5° grid. ECMWF is the only model with reforecasts (11 members) generated twice a week (Mondays and Thursdays) on the fly.

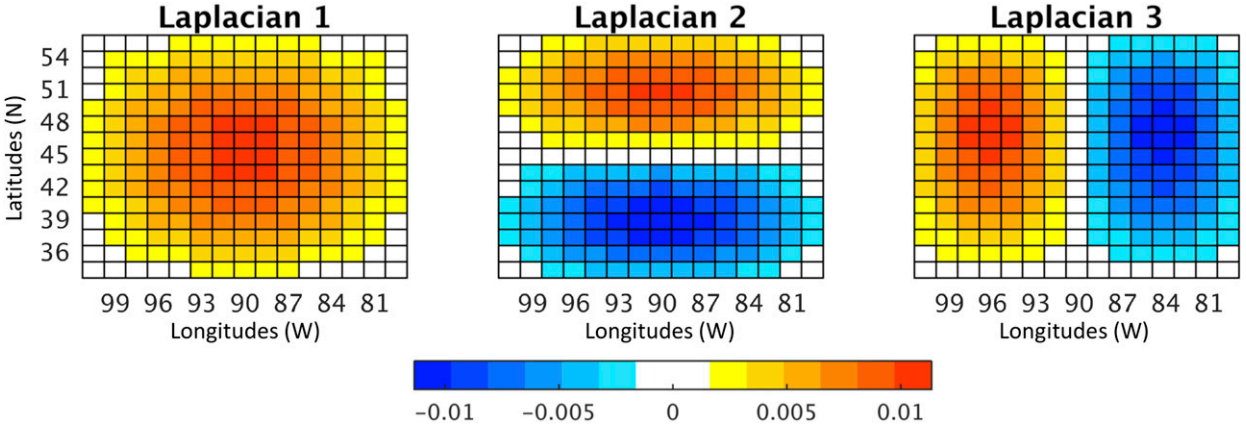

FIG. 1. First three Laplacians at 45°N, 90°W computed on a geographical box of 15 grid points in latitude and longitude.

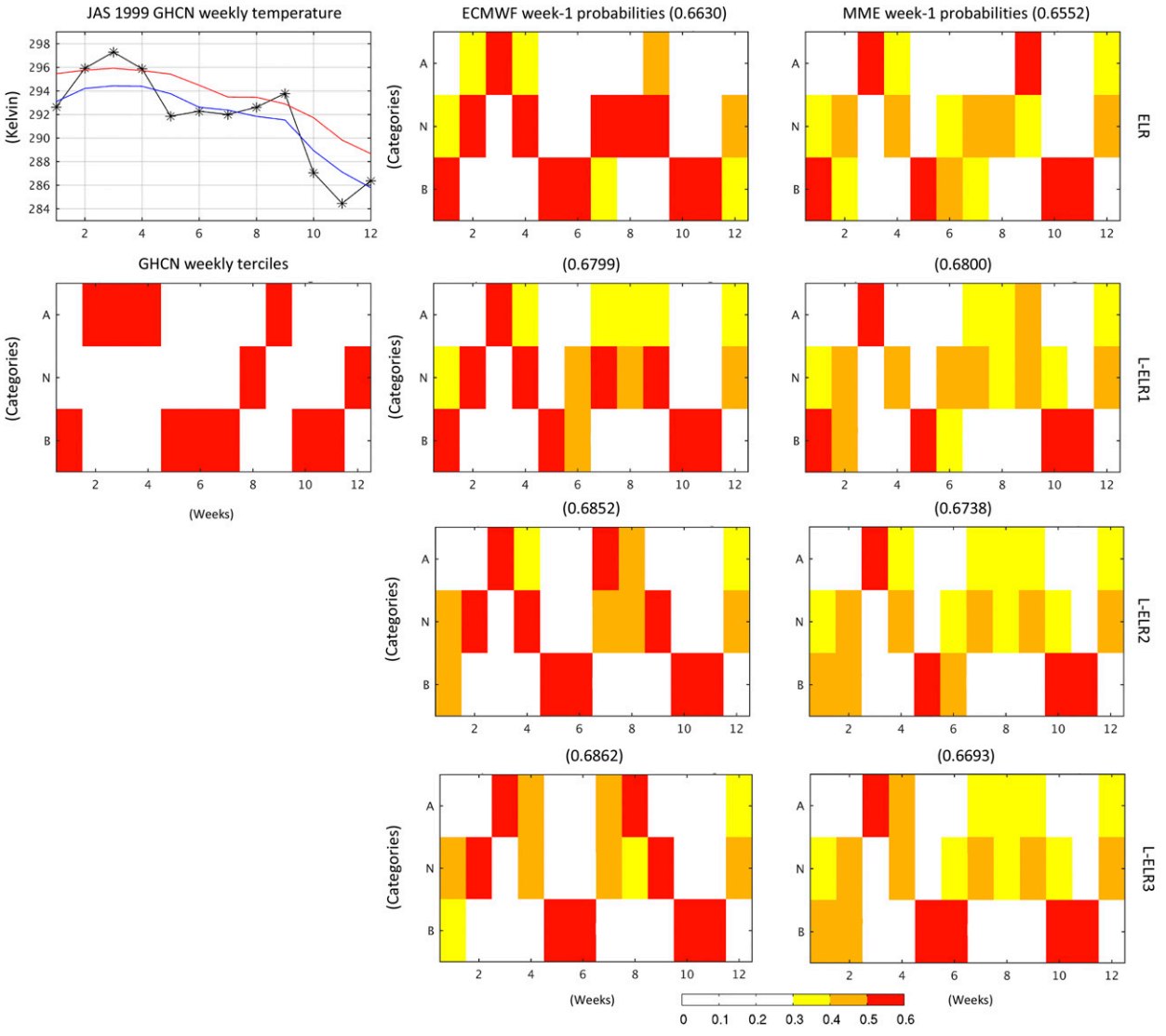

FIG. 2. Point statistics at 45°N, 90°W showing (top left) the mean GHCN surface temperatures for each week of JAS 1999 (*x* axis, i.e., from 7 Jul to 29 Sep), terciles (low and high in blue and red, respectively) and (second row, left) corresponding GHCN weekly tercile probabilities for above normal (A), normal (N), and below normal (B). Forcasted weekly tercile probabilities are shown for (center) ECMWF and (right) the multimodel ensemble (MME) of ECMWF, NCEP, and CMA hindcasts, which are pooled together with equal weighting. Mean RPS is indicated in parentheses for each forecasts.

NCEP and CMA reforecasts are generated 4 times daily from the same fixed version of their respective models. Weekly surface temperature averages from ECMWF reforecasts generated for Thusday starts in 2016 (comprising model cycles CY41R1, CY41R2, and CY43R1) are used alongside corresponding NCEP and CMA 4-member daily reforecasts, all available from 1999 to 2010, which is the period used in our study. There are thus 132 forecasts for December–February (DJF) and 144 for June–August (JJA) for each model (12 starts over 11 and 12 years, respectively). These three EPS are chosen among other S2S models because their archived reforecasts allow to design a multimodel ensemble based on exactly the same issuance dates across models, similarly to the probabilistic skill analysis of precipitation forecasts from Vigaud et al. (2017a), based on the same three models subset. Since unequal weighting is not significantly better than equal weighting in low sample size and low skill cases (DelSole et al. 2013) such as for submonthly reforecasts, forecasted probabilities from the individual models are averaged to form MME temperature tercile forecasts, whose skill is assessed over North America for winter (DJF) and summer (JJA) starts.

NOAA CPC Global Historical Climatology Network (GHCN) Climate Anomaly Monitoring System (CAMS)

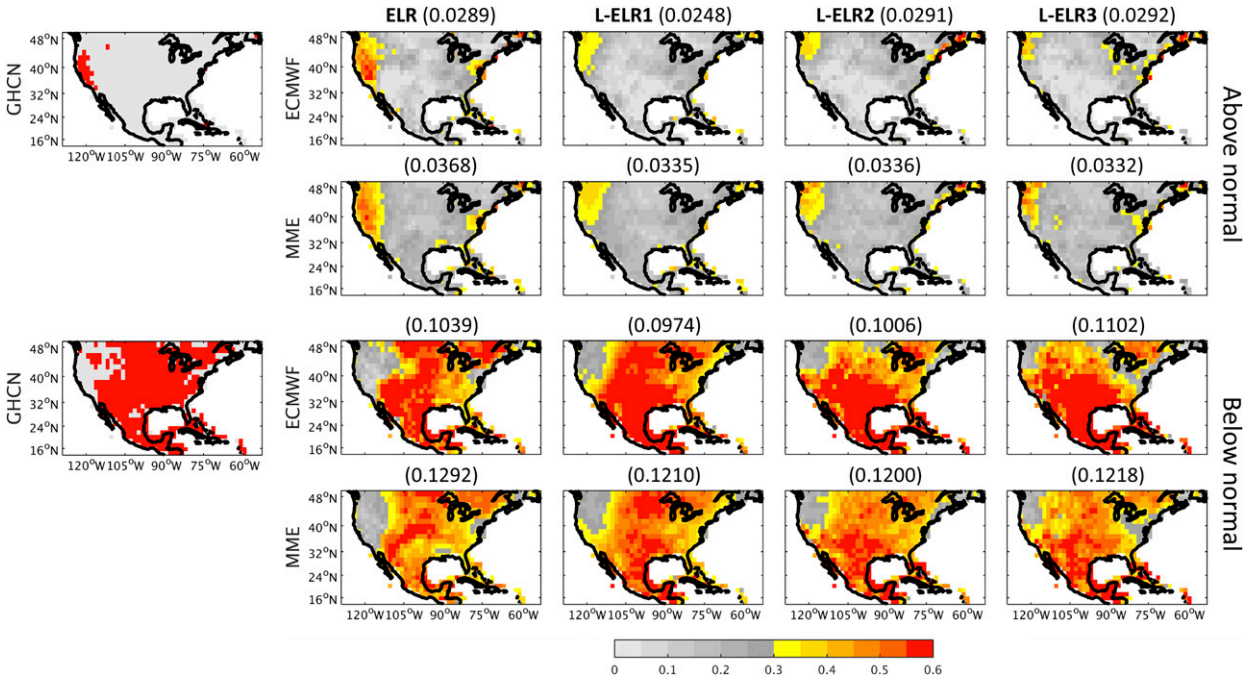

FIG. 3. (left) Observed GHCN above- and below-normal temperature tercile probabilities for 7 Jul 1999 start, together with those forecasted by ECMWF and the multimodel ensemble (MME) of ECMWF, NCEP and CMA models from ELR and L-ELR1–3 forecasts. Mean Brier score averages over the whole continental domain are indicated in parentheses for each forecast.

(Fan and van den Dool 2004, 2008) daily surface temperature estimates available from 1948 to present on a 0.5° grid, are averaged onto the common 1.5° grid of forecasts archived in the S2S database and used as observational data to calibrate and verify the reforecasts over 1999–2010.

b. *Extended logistic regression model*

The methodology is similar to ELR employed in Vigaud et al. (2017a) from which the text is derived with minor modifications as follows in this paragraph. Logistic regression is well suited to probability forecasting, and an additional explanatory variable $g(q)$ can be used to produce the probability p of nonexceedance of the quantile q:

$$\ln\left(\frac{p}{1-p}\right) = f(\overline{x_{\text{ens}}}) + g(q), \tag{1}$$

where $f = b_0 + b_1\overline{x_{\text{ens}}}$ and $g = b_2 q$. Cumulative probabilities computed from Eq. (1) for smaller predictand thresholds cannot exceed those for larger thresholds (Vigaud et al. 2017a), yielding logically consistent sets of forecasts (Wilks and Hamill 2007; Wilks 2009). ELR is computed for the 33rd and 67th temperature percentiles to produce tercile probabilities (ELR forecasts).

Observed climatological weekly tercile categories derived from GHCN weekly temperatures are defined based on 3-week windows that include the forecast target week and one week on either side, separately at each grid point for each start in DJF (8 December–25 February *Thursday* start dates) and JJA (2 June–25 August *Thursday* start dates), and each lead (from weeks 1 to week 4) following a leave-one year-out approach (i.e., using the 30 and 33 weeks from the remaining 10 and 11 years for DJF and JJA starts, respectively). ELR parameters are estimated for each model, grid point, start, and lead separately, using all years except the one being forecast, to predict terciles probabilities for the left-out year (validation set) that are averaged across models with equal weights to produce a MME of individual forecast probabilities (MME forecasts).

c. *Spatial pattern correction*

The Laplacian operator Δ in spherical coordinates λ and ϕ (longitude and latitude, respectively) is

$$\Delta f = \frac{1}{\cos^2\phi}\frac{\partial^2 f}{\partial\lambda^2} + \frac{1}{\cos\phi}\frac{\partial}{\partial\phi}\left(\cos\phi\frac{\partial f}{\partial\phi}\right). \tag{2}$$

The finite-difference approximation of Δ using a five-point stencil is

$$(\Delta f)_{i,j} = \frac{1}{\cos^2\phi_i}\left(\frac{f_{i,j+1}-f_{i,j}}{dx_{j+1}} - \frac{f_{i,j}-f_{i,j-1}}{dx_j}\right) + \frac{2}{dy_{i+1}+dy_i}\left(\frac{f_{i+1,j}-f_{i,j}}{dy_{i+1}} - \frac{f_{i,j}-f_{i-1,j}}{dy_i}\right), \tag{3}$$

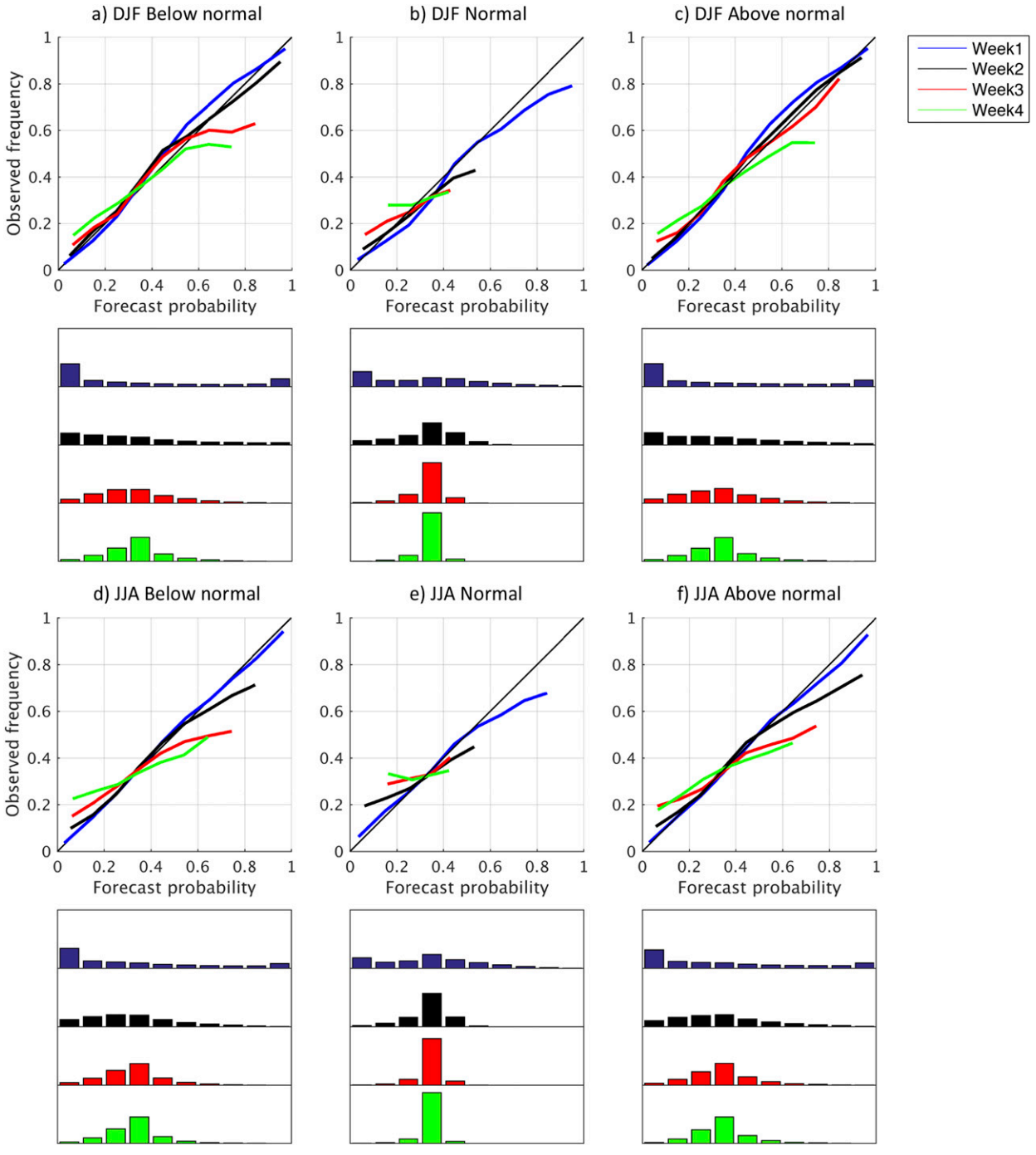

FIG. 4. Reliability diagrams for all three categories (below normal, normal, and above normal) from ECMWF ELR forecasts, with (a)–(c) DJF and (d)–(f) JJA starts, from week-1 to week-4 leads in different colors. Forecasted frequencies of issuance are shown as bins centered under the respective tercile category diagram. Forecast probabilities are plotted from 0 to 1 on the same x axis and from 0% to 100% on the y axis, and only the bins with more than 1% of all forecasts are plotted in each category. Results are computed for grid points of continental North America between 20° and 50°N latitudes.

where

$$dx_i \equiv \lambda_i - \lambda_{i-1} \quad \text{and} \quad dy_i \equiv \frac{2(\phi_i - \phi_{i-1})}{\cos\phi_i + \cos\phi_{i-1}}. \tag{4}$$

For each grid point of the North American domain, the matrix representation of Eq. (3) with Dirichlet boundary conditions is formed for the 15 × 15 grid point (e.g., 22.5° × 22.5°) box centered on that grid point. The size

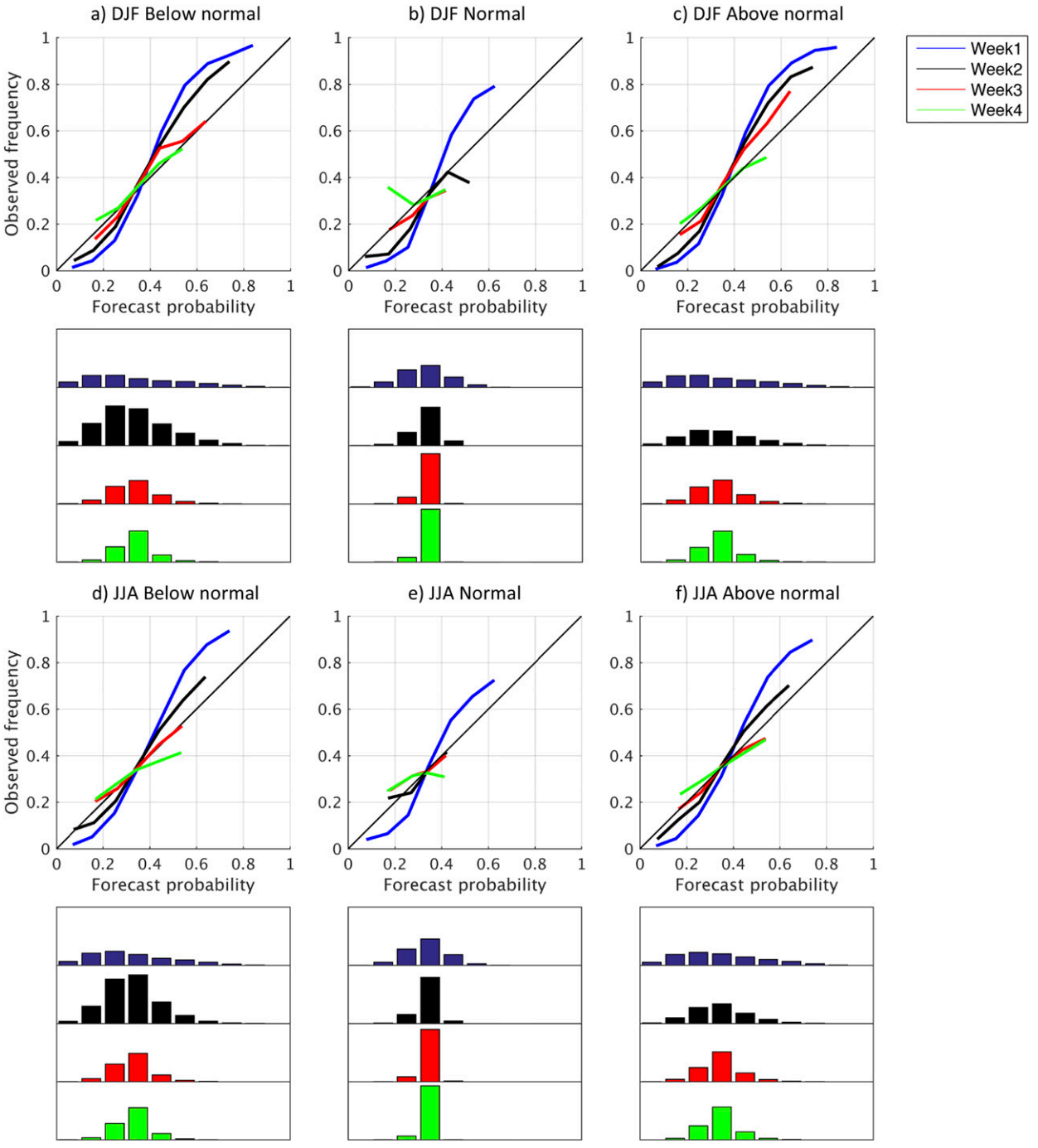

FIG. 5. As in Fig. 4, but for the multimodel ensemble (MME) of ECMWF, NCEP, and CMA ELR forecasts with starts in (a)–(c) DJF and (d)–(f) JJA.

of this box is consistent with meteorological synoptic scales such as those of midlatitude depressions for instance (thousands of kilometers), and well suited for gridpoint computations over North America. Similar results are obtained using slightly bigger or smaller boxes (not shown). The eigenvectors of this 225×225 matrix are then computed. These differ from those in DelSole and Tippett (2015) since they are computed in subdomains centered on the grid point being predicted, and they satisfy an explicit Dirichlet boundary condition.

For each model, grid point, start, and lead, forecasts are next projected with area weighting as in DelSole

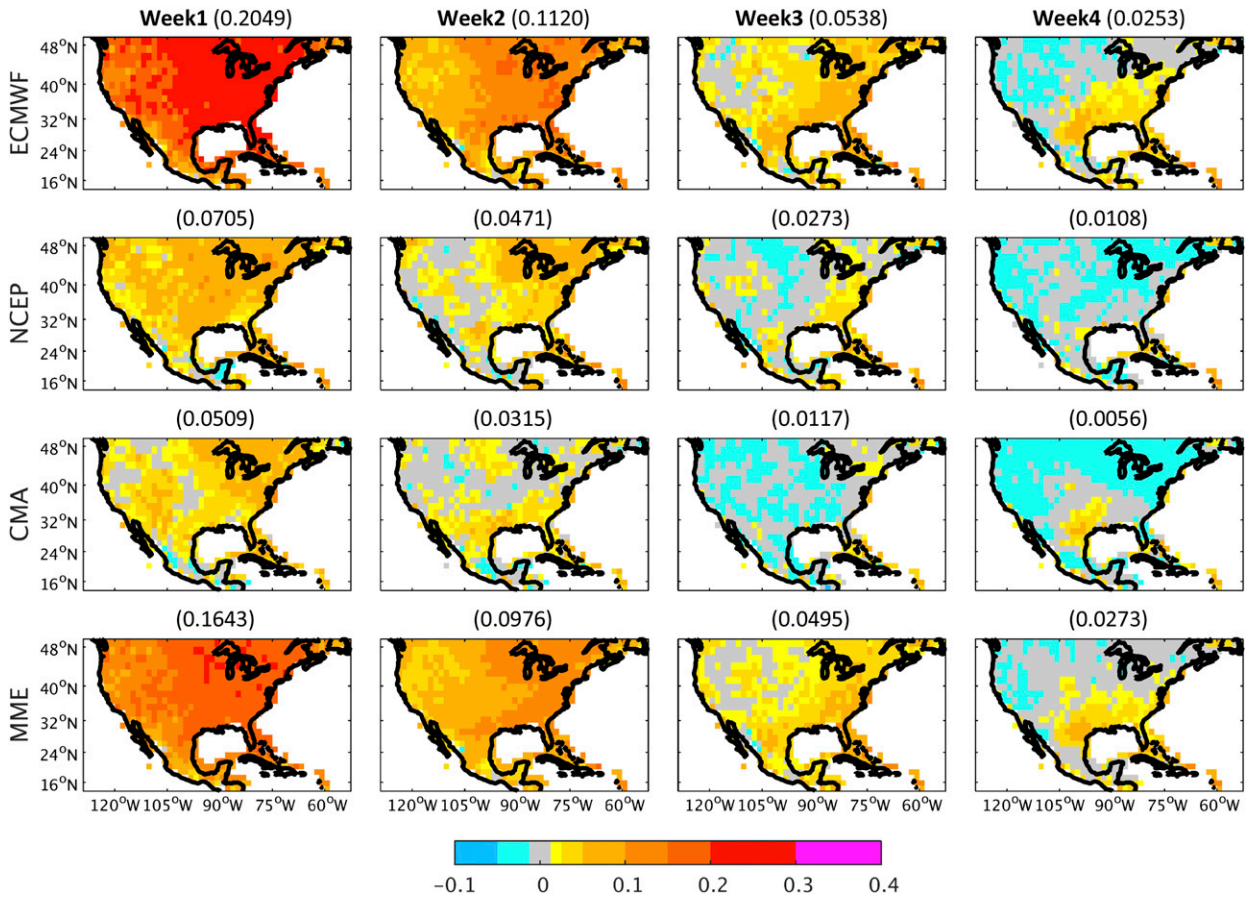

FIG. 6. Ranked probability skill scores (RPSS) for ECMWF, NCEP, and CMA temperature tercile ELR forecasts as well as their multimodel ensemble (MME) for DJF starts (rows) and different columns leads from 1 to 4 weeks (columns). Mean RPSS is indicated in parentheses for each forecast.

and Tippett (2015) onto the first three Laplacian eigenfunctions shown in Fig. 1. The first Laplacian eigenfunction represents a spatial average, while the second and third correspond to meridional and zonal gradients, respectively. The resulting amplitudes are then used as new predictors (Lap) in the ELR model, such that Eq. (1) reads:

$$\ln\left(\frac{p}{1-p}\right) = f(\mathrm{Lap}) + g(q), \tag{5}$$

where $f = b_0 + \sum_{i=1}^{n} b_i \mathrm{Lap}_i$ and $g = b_{n+1}q$ with Lap_i corresponding to the projection of the ensemble mean temperature on the ith Laplacian eigenvector. ELR models based on n eigenvectors to produce tercile probabilities will be referred to as L-ELRn forecasts, for n = 1–3.

d. Regression model setup

Weekly terciles are first defined under cross validation, as shown in Fig. 2 (left column) for GHCN observations in JAS 1999 at a grid point (45°N, 90°W). For each model, regression parameters are fitted separately at each grid point, lead and calendar start date to form weekly temperature tercile forecasts, as shown for week 1 from ECMWF weekly starts in Fig. 2 (middle column). ECMWF category forecasts display highest weekly probabilities consistent with observed terciles and are more skillful than those from NCEP and CMA (not shown). Similarly to Vigaud et al. (2017a), the three forecasts are averaged with equal weights to produce MME forecasts shown in Fig. 2 (right column). MME has the same or slightly lower RPS values, indicating a moderate increase in skill.

Probability maps from forecasts initialized 7July 1999 (Fig. 3) display highest probabilities consistent with GHCN, MME forecasts being more skillful than ECMWF with comparable skill levels for ELR and L-ELR forecasts.

e. Skill metrics and significance testing

Reliability diagrams are computed by pooling all land grid points over continental North America between 20°

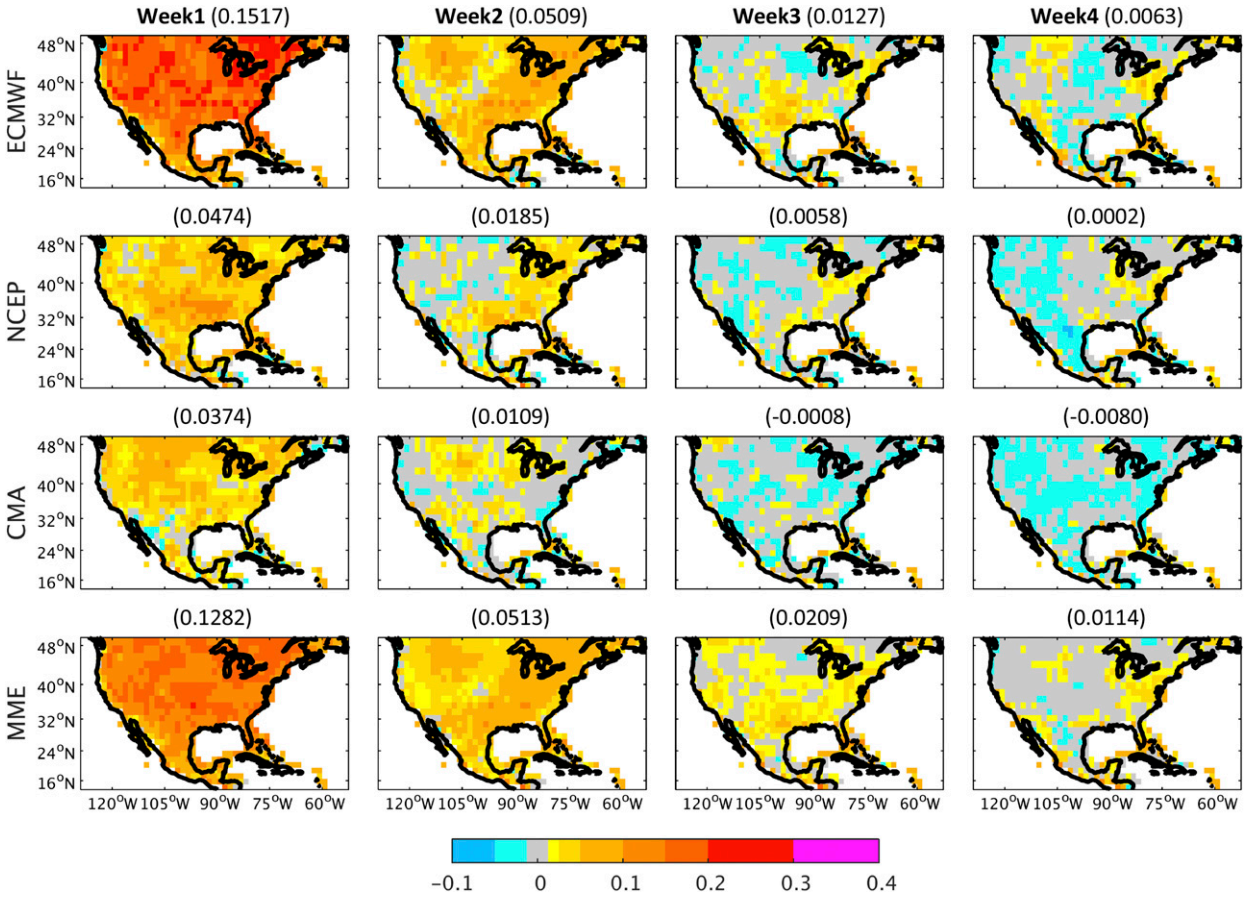

FIG. 7. As in Fig. 6, but for starts during the JJA season.

and 50°N to evaluate the reliability, but also resolution as well as sharpness (Wilks 1995; Hamill 1997), of ELR and L-ELR tercile category temperature forecasts. Spatial information is provided by maps of ranked probability skill scores (RPSSs; Epstein 1969; Murphy 1969, 1971; Weigel et al. 2007) that quantify the extent to which calibrated predictions are improved in comparison to climatological frequencies. RPSS is one of the most commonly used strictly proper skill scores (Daan 1985; Wilks 1995; Weigel et al. 2007), and its values tend to be small; a reliable deterministic forecast with correlation r will have a RPSS of approximately $1 - \sqrt{1 - r^2}$ [i.e., a RPSS value of 0.1 corresponds to a correlation near 0.44; Tippett et al. (2010)].

Monte Carlo simulations based on large numbers of random forecasts samples (i.e., 100 000) drawn from all forecasts with DJF and JJA starts are used separately to assess the significant area averages RPSS during specific ENSO conditions and MJO phases, which are, respectively, compared to the 90th percentile RPSS derived from all winter and summer starts. Monte Carlo simulations are also used to assess the significance of the correlations of area averages of weekly MME RPSSs with the observed Niño-3.4 index (Barnston et al. 1997) and real-time multivariate MJO (RMM) indices (Wheeler and Hendon 2004), and of these indices with observed weekly rainfall.

3. Results

a. Baseline ELR weekly forecasts

Reliability diagrams for weekly ECMWF ELR forecasts with DJF and JJA starts (Fig. 4) show good reliability and resolution for week 1 in both seasons, as indicated by the blue lines near the diagonal and away from the 0.33 horizontal line (not plotted), respectively. Histograms of forecast probabilities spread across all bins in week 1 and indicate high sharpness, while forecast frequencies are clustered toward equal odds as lead increases. From week 2, reliability and resolution drop, with more skill in winter than summer. NCEP and CMA forecasts exhibit similar results but are overall less skillful (not shown).

Greater slopes for the MME (Fig. 5) reflect underconfidence and lack of resolution at most leads other

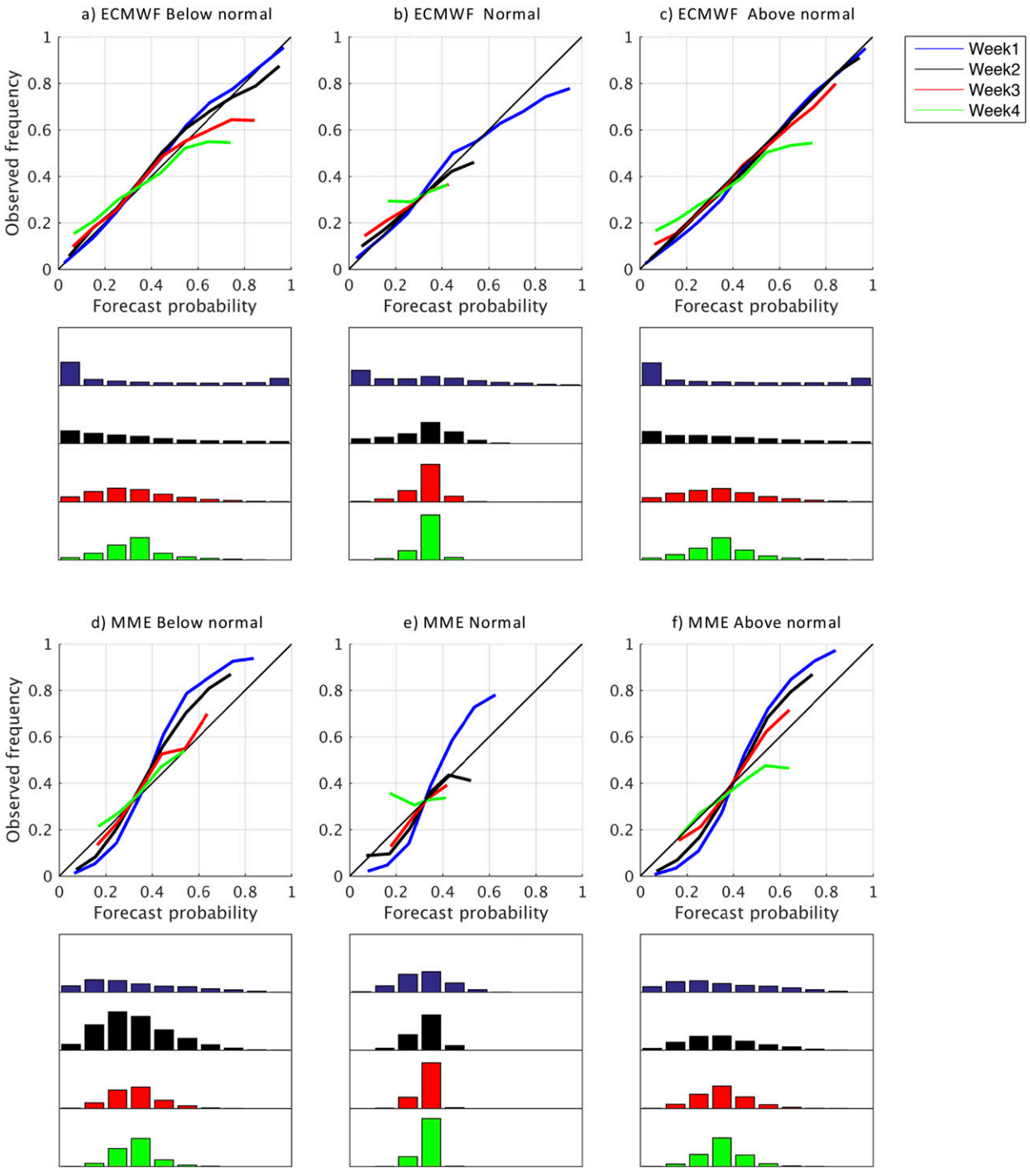

FIG. 8. As in Fig. 4, but for (a)–(c) ECMWF and the (d)–(f) multimodel ensemble (MME) of ECMWF, NCEP, and CMA L-ELR1 forecasts with DJF starts.

than lead 4, with high sharpness but lower than for ECMWF at all leads. Skill also decreases with increasing leads, week-4 MME forecasts showing only small deviations from equal odds, and from winter to summer.

Positive RPSS values for week-1 forecasts from individual models and their MME starting in DJF (Fig. 6) are maximum east of 100°W, where largest RPSS remains with half the magnitude in week 2, when ECMWF is the most skillful model and CMA exhibits lowest skill.

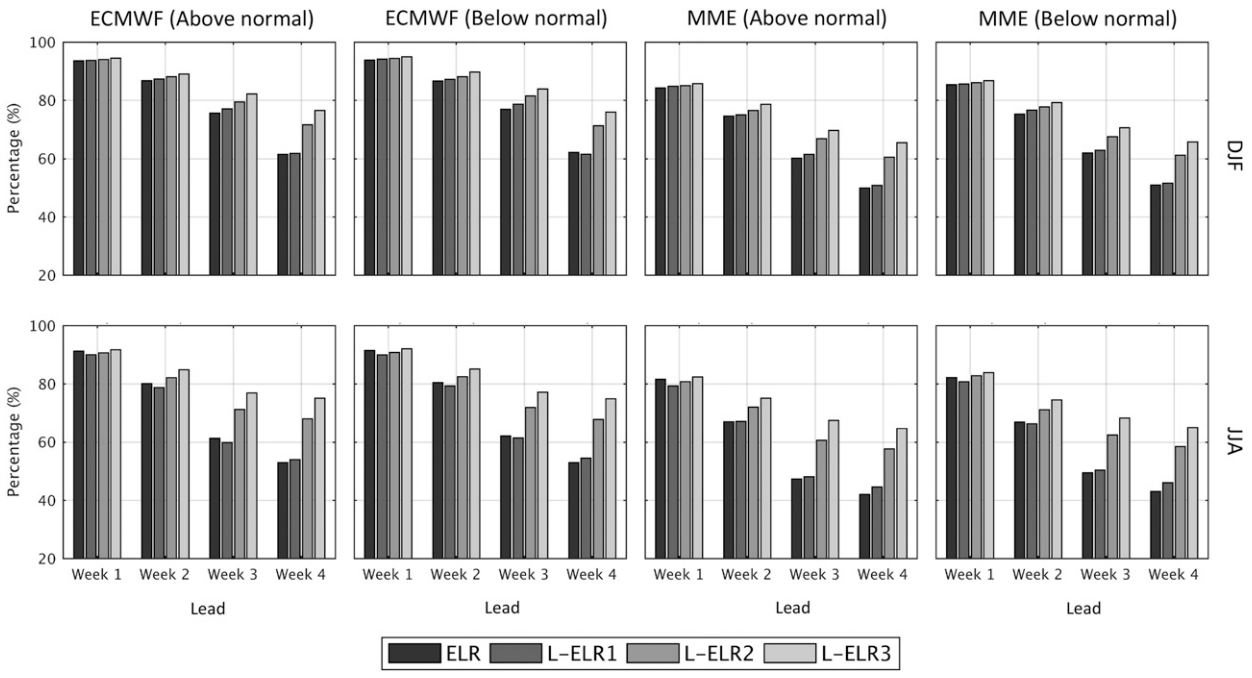

FIG. 9. Percentages of forecasts outside for the fourth bin (0.33) for week-1–4 forecasts from (left) ECMWF and (right) the MME for the above and below-normal categories averaged over continental North America between 20° and 50°N latitudes, for temperature tercile forecasts with (top) DJF and (bottom) JJA starts.

Near-zero or negative values are found everywhere at higher leads, except for ECMWF in week 3 that is still skillful over the eastern United States, and for the southern U.S.–Gulf of Mexico (GoM) regions where skill holds in ECMWF and CMA. Multimodel combination does not result in marked RPSS increase for week-1 to week-3 forecasts compared to the most skillful ECMWF model. In week 4, however, multimodel ensembling damps negative skill values in individual forecasts and reflects maximum RPSS values for ECMWF and CMA forecasts over the southern United States–GoM, where skill is highest. Forecast skill is higher in winter than summer (Fig. 7), agreeing with Figs. 4 and 5. For JJA starts, skill levels in week 1 are comparable across models and drop from week 2 with near-zero or negative RPSS values prevailing from week 3, except over the northeast United States in ECMWF, NCEP, and the MME showing also maximum RPSS over Mexico and the GoM.

Over North America, ECMWF generally produces the most reliable and skillful temperature tercile forecasts of all three EPSs from week-1 to week-4 leads. The relatively poor performance of NCEP and CMA models past week 2 translates into limited or no skill improvement from multimodel ensembling for both winter and summer starts. Including more models available from both the S2S and SubX (Pegion et al. 2019) databases alongside differential weighting schemes could potentially help improving skill, but this needs to be tested in further studies.

b. *Skill improvements with spatial pattern correction*

Reliability diagrams for ECMWF and MME L-ELR1 forecasts, using one Laplacian eigenfunction (e.g., the spatial average of ensemble mean temperature in Fig. 1) as predictor instead of the gridpoint mean temperature, exhibit comparable reliability for DJF starts in Fig. 8 than those from ELR (Figs. 4 and 5). Figure 9 shows spatial averages over North America between 20° and 50°N latitudes of the percentages of ELR forecasts different from climatology in Figs. 4 and 5, which is an indication of sharpness, alongside those from L-ELR1 (Fig. 8) to L-ELR3 (not shown) forecasts. Sharpness decreases with lead for all forecasts and is comparable between L-ELR1 and ELR at week 1, but increases for L-ELR1 from weeks 2 and 3 in DJF and JJA, respectively. Noteworthy, L-ELR2–3 forecasts are increasingly sharper than ELR and L-ELR1 with increasing leads, reflecting overconfidence.

RPSS for L-ELR1–3 forecasts with DJF and JJA starts (Figs. 10–13) have comparable structures to those from ELR with less negative values for the MME than ECMWF. Overall, skill improvement by multimodel ensembling is limited compared to the best

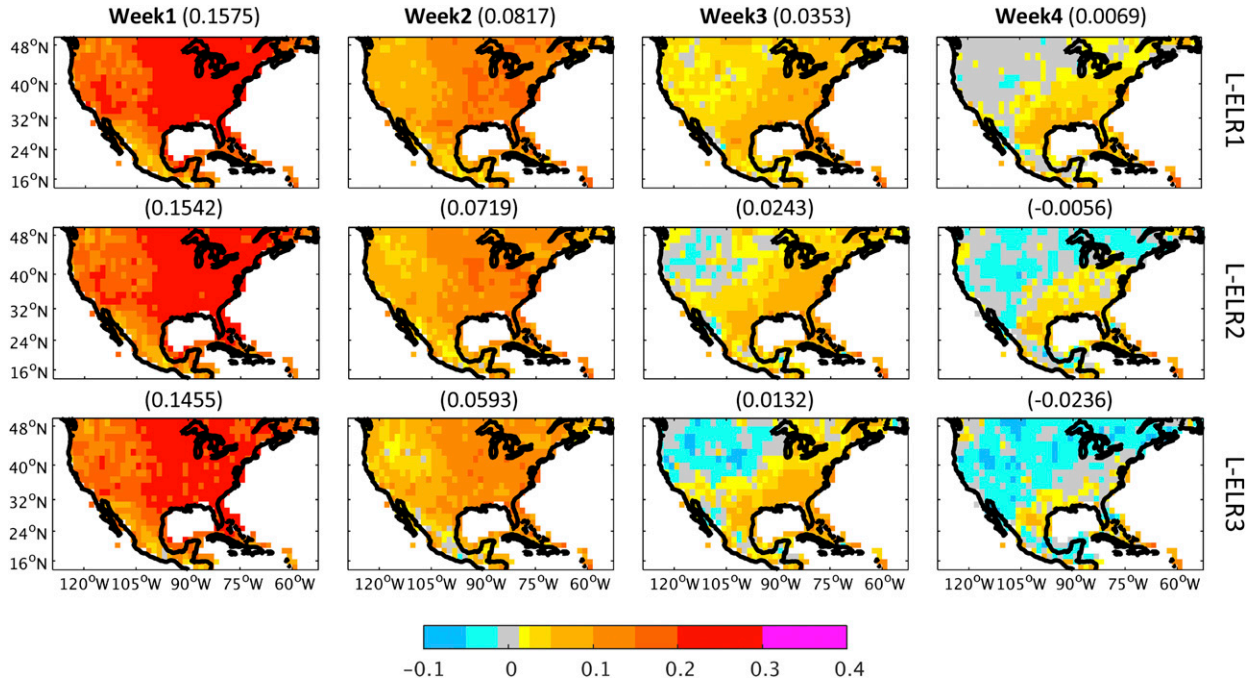

FIG. 10. RPSS for ECMWF L-ELR1–3 temperature tercile forecasts for DJF starts (rows) and different leads from 1 to 4 weeks (columns). Mean RPSS is indicated in parentheses for each forecast.

ECMWF model, as noted for baseline ELR forecasts. Higher RPSS for ECMWF and MME L-ELR1 forecasts compared to ELR from week 3 translates into highest skill over North America (Fig. 14) that contrasts with comparable skill levels at shorter leads. RPSS values drop when adding more predictors in L-ELR2–3 and, together with increased sharpness (Fig. 9), suggest overconfidence and reduced reliability. This overconfidence

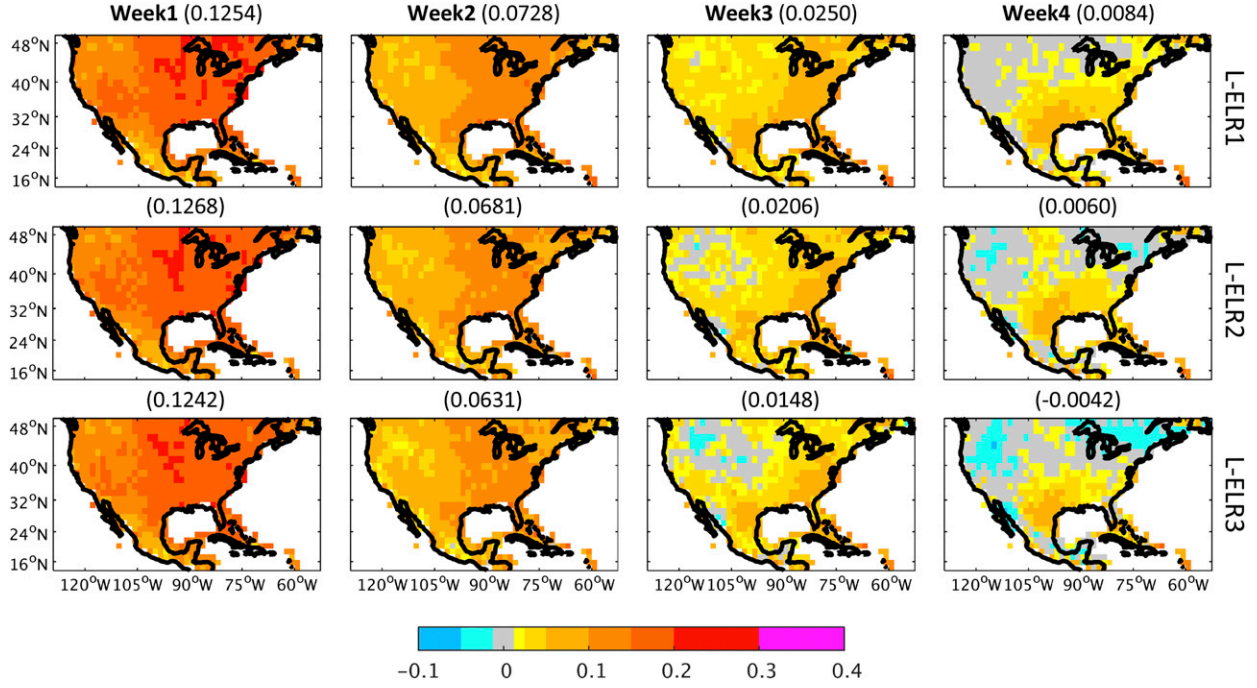

FIG. 11. As in Fig. 10, but for the multimodel ensemble (MME) of ECMWF, NCEP, and CMA.

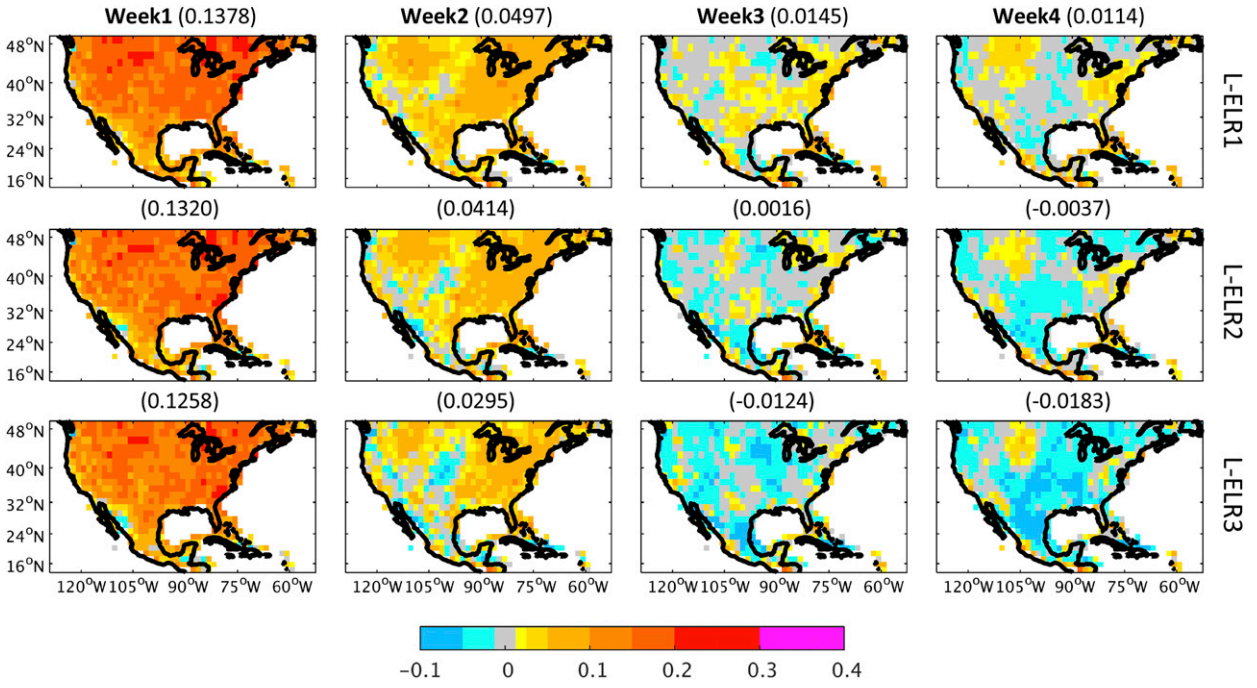

FIG. 12. As in Fig. 10, but for starts during the JJA season.

can be related to the sensitivity of regression methods to sample variability, which increases with the number of coefficients being estimated and can be reduced by increasing sample size (Tippett et al. 2014). The short length of reforecasts used for training at each start date (three reforecasts over 10 and 11 years for DJF and JJA starts, respectively) does not allow to significantly satisfy the rule of thumb of having approximately

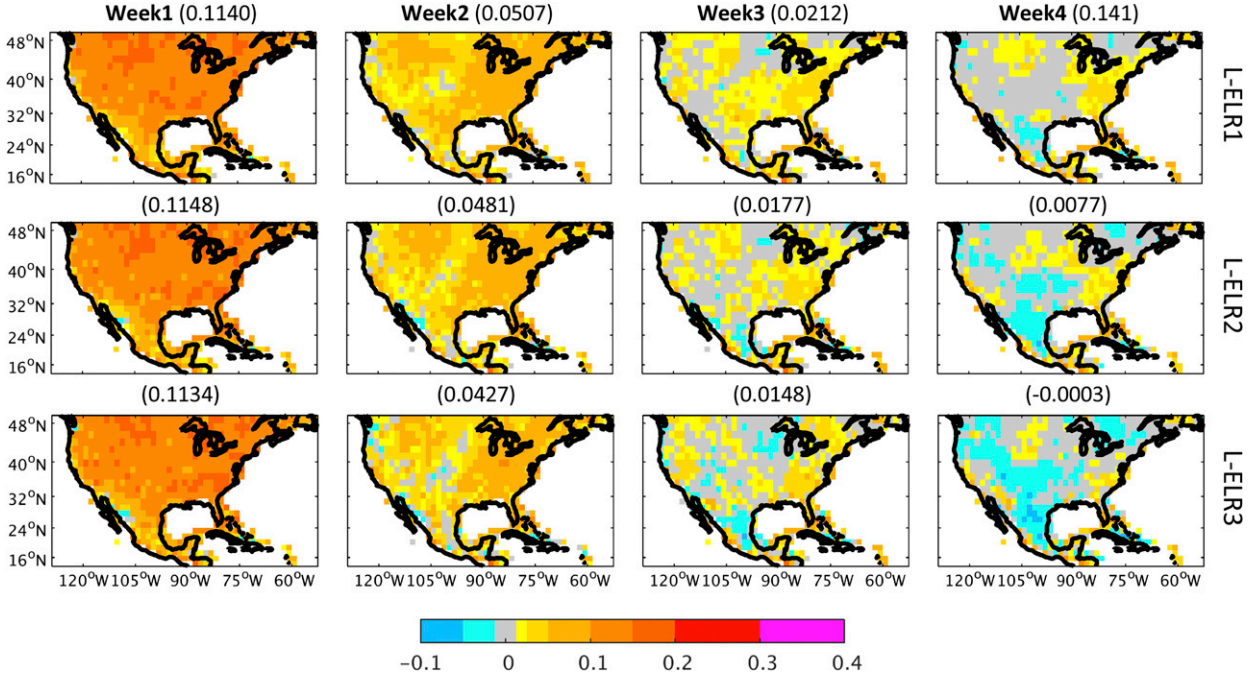

FIG. 13. As in Fig. 12, but for the multimodel ensemble (MME) of ECMWF, NCEP, and CMA.

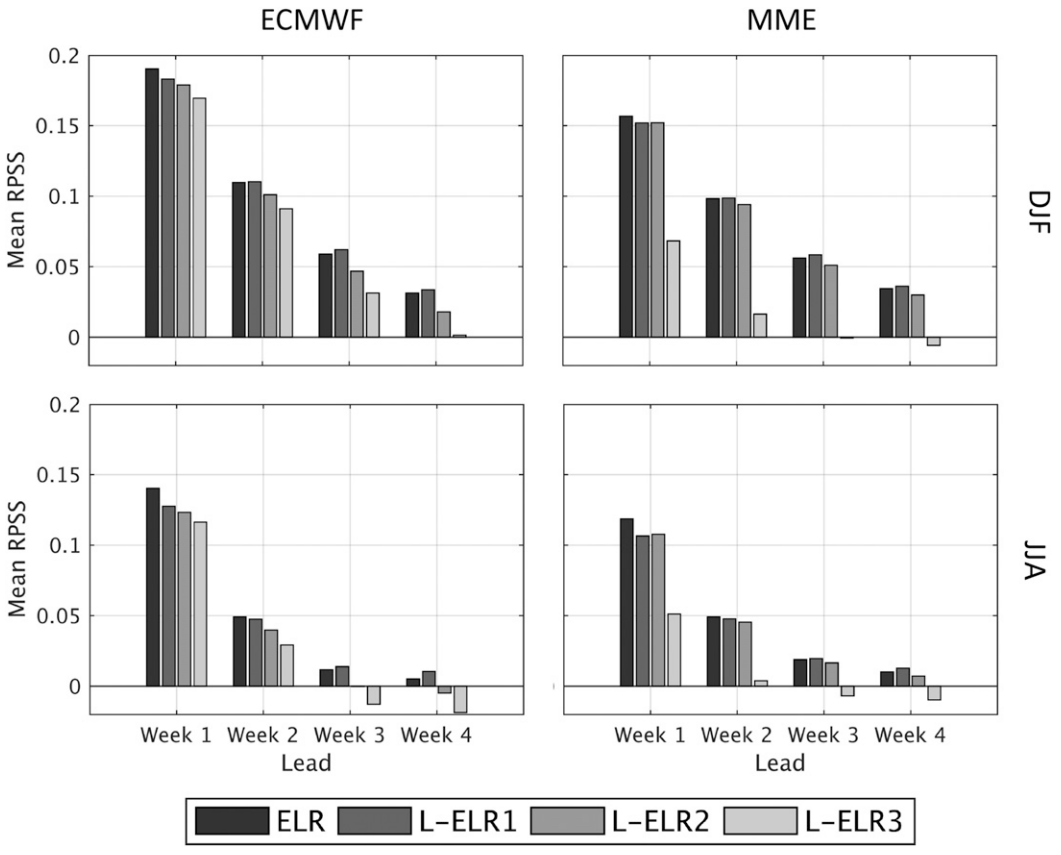

FIG. 14. Mean weekly RPSS averaged over continental North America between 20° and 50°N latitudes, for week-1–4 ELR and L-ELR1–3 temperature tercile forecasts from (left) ECMWF and (right) the multimodel ensemble (MME) with (top) DJF and (bottom) JJA starts.

10 samples per explanatory variables, beyond two predictors.

c. *Skill relationships to ENSO and the MJO*

Significant correlations between weekly GHCN temperature estimates and the Niño-3.4 (Barnston et al. 1997) and RMM indices (Wheeler and Hendon 2004) in Fig. 15 (top panels) suggest forecast skill relationships to both large-scale signals, particularly in winter. This is further confirmed by resemblances between both Niño-3.4 and RMM1 correlations and the spatial correlation patterns of the first principal components (PCs) obtained by applying a principal component analysis (PCA) to weekly MME RPSS values (i.e., the mean is not removed) in Fig. 15 bottom panels. In winter, these PCs are highly correlated to mean RPSS (above 0.9) and account for a significant part of total variance from week-1 to week-4 leads (near and above 30%). Maximum PC1 loadings over the east United States at all leads coincide with anticorrelations between weekly temperatures and Niño-3.4, alongside similarities to RMM1 positive correlation pattern, that are consistent with Table 2. In summer, PC1 is also highly correlated to mean RPSS but explains lower amount of variance beyond week 1 (below 15%) and its maximum loadings correspond less well with pattern correlations of weekly temperatures, except over the southwest and southeast United States, where positive correlations are also typical of Niño-3.4.

Weekly MME RPSS values and above-normal forecast probabilities averaged over North America between 20° and 50°N, and stratified by ENSO conditions (El Niño and La Niña when Niño-3.4 is greater and lower than 0.5, respectively, and neutral conditions otherwise) and distinct MJO phases, are shown for forecasts with DJF starts in Fig. 16. The small reforecast sample contains no strong El Niño event and no significant skill relationship is found with ENSO phases in winter, except for La Niña phases in week 3 and 4, but skill remains low. This is reflected by barely significant above-normal probabilities at both leads, that can be related to maximum skill for DJF starts over the southeast United States (Figs. 6, 10, and 15), where warmer conditions generally prevail in winter for cold

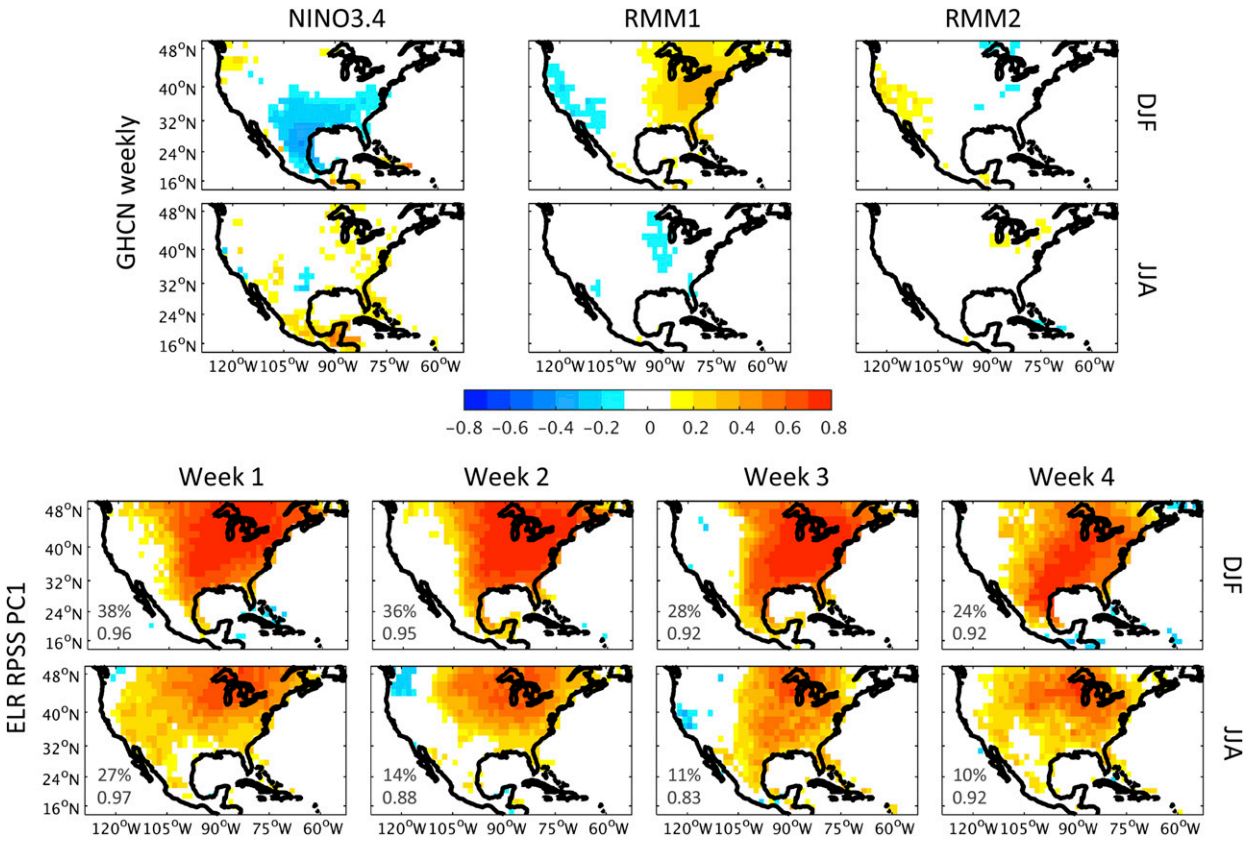

FIG. 15. The upper panels show spatial correlation patterns of (top) DJF and (bottom) JJA GHCN weekly temperatures and (left) observed weekly Niño-3.4 index (Barnston et al. 1997) and (center),(right) RMM indices (Wheeler and Hendon 2004). The lower panels show weekly MME leading RPSS PC1 from ELR forecasts with starts in (top) DJF and (bottom) JJA. Only correlations significant at the 0.05 level using Monte Carlo simulations are plotted. The fraction of total variance explained by each PC is indicated in the different panels (%) as well as their correlations to spatially averaged RPSSs.

ENSO phases (Smith and Sardeshmukh 2000), consistent with Fig. 15 and Table 2, also suggesting that above-normal probabilities below 33% could be related to cooling over the Gulf coast through El Niño-induced jet modulations (Ropelewski and Halpert 1986).

Higher mean RPSS across MJO phases than for ENSO might indicate more predictability from the MJO. Winter skill is increased for MJO phase 3 up to week 3 and phase 6 up to week 2 (Fig. 16), when convection is enhanced over the Indian Ocean and western Pacific, respectively. This is consistent with North

TABLE 2. Correlations between weekly MME RPSSs averaged over North America between 20° and 50°N in DJF and JJA, as well as its leading principal components (PC1, in parentheses), and the observed Niño-3.4 index (second column), MJO measured by the RMM1 (third column), and RMM2 (fourth column) indices of Wheeler and Hendon (2004), and their best linear combination (fifth column). Scores significant at the 0.05 level of significance using Monte Carlo simulations are indicated with *.

Mean (PC1) RPSS	Niño-3.4	RMM1	RMM2	MJO
DJF week 1	−0.13 (−0.17)	0.16* (0.21*)	0.02 (−0.07)	−0.15 (−0.19*)
DJF week 2	−0.08 (−0.11)	0.15 (0.17*)	−0.16 (−0.22*)	−0.17 (−0.24*)
DJF week 3	−0.09 (−0.16)	−0.09 (−0.07)	−0.17 (−0.23)	−0.19* (−0.25*)
DJF week 4	−0.20* (−0.39*)	−0.12 (−0.06)	−0.16* (−0.14)	−0.21* (−0.16)
JJA week 1	0.24* (0.21*)	−0.1 (−0.13)	−0.06 (−0.03)	0.02 (−0.09*)
JJA week 2	0.21* (0.17*)	0.02 (−0.01)	0 (0.05)	0.02 (−0.04)
JJA week 3	0.31* (0.27*)	−0.12 (−0.14)	0.01 (0.06)	0.12 (0.15)
JJA week 4	0.25* (0.22*)	−0.04 (−0.09)	−0.09 (−0.07)	0.1 (0.13)

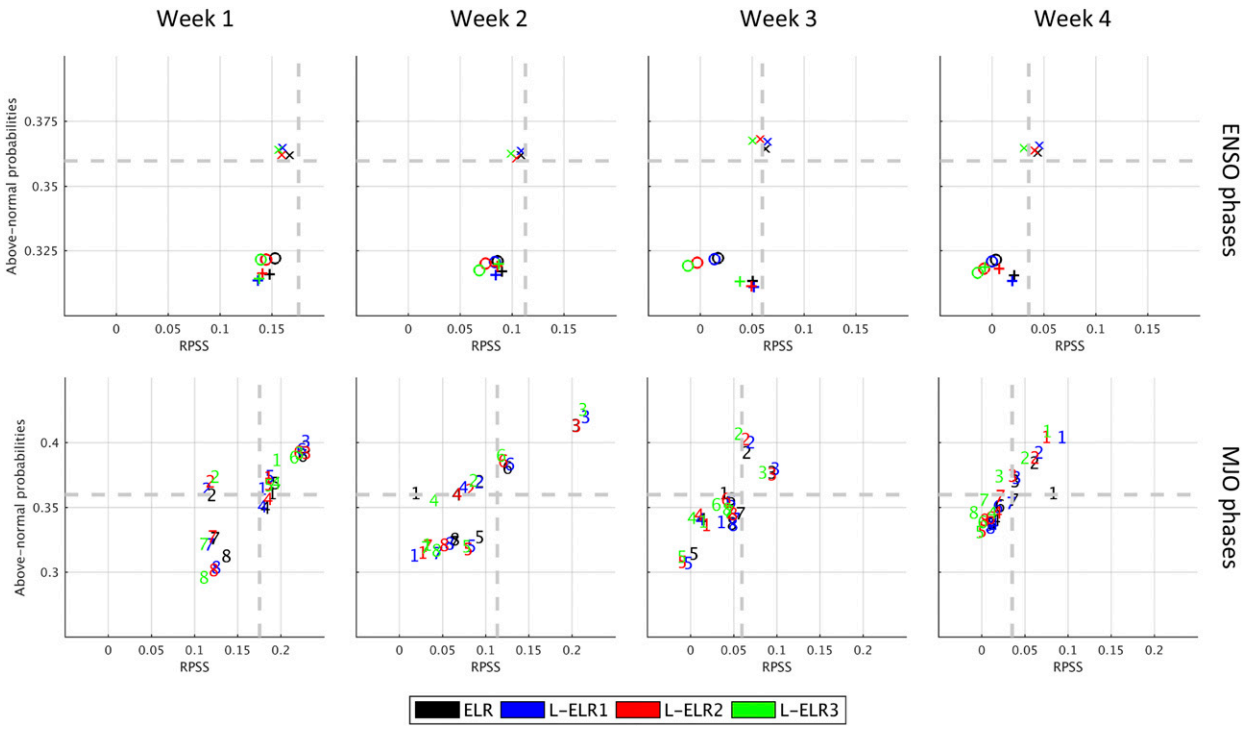

FIG. 16. (top) Mean weekly RPSS averaged over continental North America between 20° and 50°N for week 1–4 MME ELR and L-ELR1–3 temperature tercile forecasts with DJF starts during observed phases of the Niño-3.4 index (Barnston et al. 1997) vs corresponding above-normal probabilities, where El Niño, neutral ENSO, and La Niña phases are indicated by +, o, and x symbols, respectively. Dashed lines correspond to a 0.1 level of significance using Monte Carlo simulations. (bottom) As in the top panels, but for MJO phases measured by RMM indices (Wheeler and Hendon 2004).

American temperatures relationship to MJO phases characterized by strong dipolar anomalies in tropical diabatic heating and convection (Lin and Brunet 2009; Yao et al. 2011; Rodney et al. 2013) associated with anomalous Rossby waves (Hoskins and Karoly 1981; Karoly 1983; Hoskins and Ambrizzi 1993) favoring extratopical teleconnections (Lin et al. 2009, 2010; Lin and Brunet 2018) and impacting winter temperatures with a precursive signal up to 2-week lead around phases 3 and 6 (Lin and Brunet 2009; Yao et al. 2011). Highest RPSSs for MJO phases 3 and 6 coincide with enhanced above-normal forecast probabilities, consistent with positive RMM1 correlations to east coast temperatures (Fig. 15) and to RPSS (Table 2), which maximum RPSS and above-normal probabilities in week 4 for phases 1–3 might be reminiscent of.

Skill levels are lower in summer across both ENSO and MJO phases (Fig. 17) compared to winter. The highest RPSS values occur during El Niño at all leads compared to neutral and La Niña phases, and are consistent with Table 2. Maximum positive correlations between Niño-3.4 and weekly temperatures over the southeast and southwest United States with parts of positive PC1 loadings over these regions (Fig. 15) suggest enhanced predictability there during warm phases of ENSO, but forecast probabilities close to climatological odds reflect weak relationships. Sustained wave teleconnections could explain maximum skill at most leads for MJO phase 3, and in week 1 for phase 8, while skill is near zero or negative otherwise after week 2, and above-normal forecast probabilities are barely significant beyond week 1.

4. Discussion and conclusions

The skill of S2S forecasts from ECMWF, NCEP, and CMA week 1–4 leads has been investigated by applying ELR to produce weekly tercile probabilities over the common 1999–2010 period. While baseline forecasts use the gridpoint ensemble mean as predictor, spatial correction is next implemented through the decomposition of the ensemble mean temperature neighboring each grid point, using locally defined Laplacian eigenfunctions (Fig. 1). Individual model probabilities are averaged to form the multimodel ensemble (MME) forecasts (Figs. 2 and 3). Over North America, weekly temperature tercile forecasts based on the gridpoint ensemble mean are characterized by high sharpness and decreasing skill

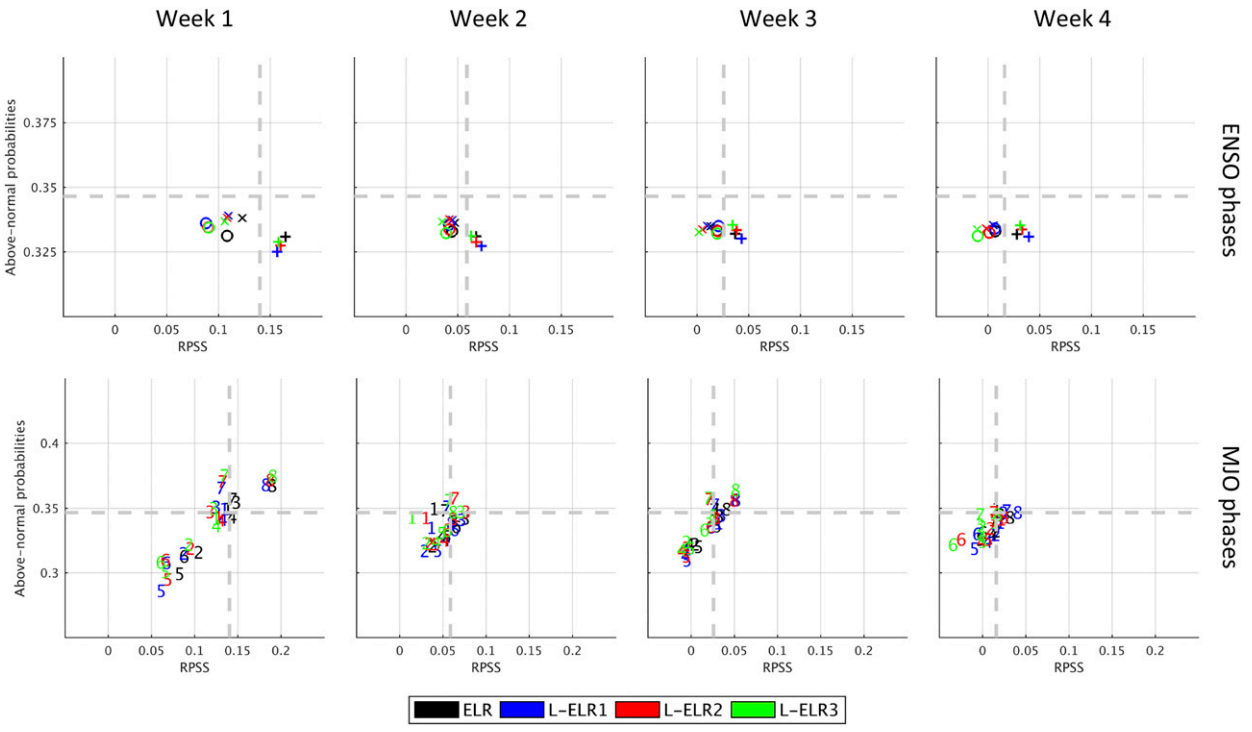

FIG. 17. As in Fig. 16, but for starts during the JJA season.

with lead times for starts in DJF and JJA, when reliability and resolution drop after weeks 1 and 2, respectively (Figs. 4 and 5), with more skillful predictions in winter than summer (Figs. 6 and 7). When using the first Laplacian eigenfunction (i.e., the spatial average of ensemble mean temperature) instead of the gridpoint ensemble mean, forecasts are characterized by comparable sharpness, resolution and reliability at weeks 1 and 2 (Figs. 8 and 9), but skill levels are slightly increased from week 3 (Figs. 10–14). Skill decreases when including more Laplacian eigenfunctions as additional explanatory variables, which can be related to the sensitivity of regressions to sample variability, suggesting that improvements are limited by the small size of reforecasts used to train the ELR model. Overall, there is no substantial skill improvement by multimodel ensembling compared to the ECMWF model for all forecasts and both seasons, even when spatial pattern corrections are applied. Including more models and multimodel ensembling approaches with unequal weighting could potentially help to improve skill, and this needs to be further studied.

Weekly temperature and skill relationships to ENSO and the MJO (Figs. 15–17 and Table 2) suggest modulations from both large-scale signals. Significant but weak skill relationships are identified in winter with La Niña at week-3 and week-4 leads (Fig. 16), potentially reflecting warm conditions for cold ENSO phases over the southeast United States, where skill is maximum in DJF (Fig. 15) and contrasts with the cold bias in seasonal model forecasts. Skill is increased for summer starts during El Niño at all leads, but remains small (Fig. 17), with associated forecasted probabilities close to climatological odds. MJO modulates skill more significantly in winter with the highest skill in both seasons up to week 3 coinciding with enhanced above-normal probabilities for MJO phase 3, when MJO-induced dipolar anomalies are known to favor extratropical teleconnections and skill could be potentially related to the predictability of Rossby waves that influence North American temperature. Such opportunities for skillful predictions could be exploited in future studies and translate into useful climate information for applications in the S2S time range.

Acknowledgments. The authors thank the anonymous reviewers for their insightful feedbacks, which helped in improving the manuscript substantially. The authors are grateful to the financial support received from the NOAA-NWS Next Generation Global Prediction System (NGGPS) Testbed Research-to-Operation (R2O) Project Award NA18NWS4680067. Computations were performed using IRI resource,s and this work is based on GHCN and S2S data archived in the IRI Data Library (IRIDL, http://iridl.ldeo.columbia.edu). S2S is a joint

initiative of the World Weather Research Programme (WWRP) and the World Climate Research Programme (WCRP). The original S2S database is hosted at ECMWF as an extension of the TIGGE database.

REFERENCES

Barnston, A., and C. Ropelewski, 1992: Prediction of ENSO episodes using canonical correlation analysis. *J. Climate*, **5**, 1316–1345, https://doi.org/10.1175/1520-0442(1992)005<1316:POEEUC>2.0.CO;2.

——, and M. Tippett, 2017: Do statistical pattern corrections improve seasonal climate predictions in the North American multimodel ensemble models? *J. Climate*, **30**, 8335–8355, https://doi.org/10.1175/JCLI-D-17-0054.1.

——, M. Chelliah, and S. Goldenberg, 1997: Documentation of a highly ENSO-related SST region in the equatorial Pacific: Research note. *Atmos.–Ocean*, **35**, 367–383, https://doi.org/10.1080/07055900.1997.9649597.

Daan, H., 1985: Sensitivity of verification scores to the classification of the predictand. *Mon. Wea. Rev.*, **113**, 1384–1392, https://doi.org/10.1175/1520-0493(1985)113<1384:SOVSTT>2.0.CO;2.

DelSole, T., and M. Tippett, 2015: Laplacian eigenfunctions for climate analysis. *J. Climate*, **28**, 7420–7436, https://doi.org/10.1175/JCLI-D-15-0049.1.

——, X. Yang, and M. Tippett, 2013: Is unequal weighting significantly better than equal weighting for multi-model forecasting? *Quart. J. Roy. Meteor. Soc.*, **139**, 176–183, https://doi.org/10.1002/qj.1961.

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

Fan, Y., and H. van den Dool, 2004: Climate prediction center global monthly soil moisture data set at 0.5° resolution for 1948 to present. *J. Geophys. Res.*, **109**, D10102, https://doi.org/10.1029/2003JD004345.

——, and ——, 2008: A global monthly land surface air temperature analysis for 1948–present. *J. Geophys. Res.*, **113**, D01103, https://doi.org/10.1029/2007JD008470.

Feddersen, H., A. Navarra, and M. N. Ward, 1999: Reduction of model systematic error by statistical correction for dynamical seasonal predictions. *J. Climate*, **12**, 1974–1989, https://doi.org/10.1175/1520-0442(1999)012<1974:ROMSEB>2.0.CO;2.

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

——, J. Whitaker, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.*, **132**, 1434–1447, https://doi.org/10.1175/1520-0493(2004)132<1434:ERIMFS>2.0.CO;2.

Hoskins, B., and D. Karoly, 1981: The steady linear response of a spherical atmosphere to thermal orographic forcing. *J. Atmos. Sci.*, **38**, 1179–1196, https://doi.org/10.1175/1520-0469(1981)038<1179:TSLROA>2.0.CO;2.

——, and T. Ambrizzi, 1993: Rossby wave propagation on a realistic longitudinally varying flow. *J. Atmos. Sci.*, **50**, 1661–1671, https://doi.org/10.1175/1520-0469(1993)050<1661:RWPOAR>2.0.CO;2.

Karoly, D., 1983: Rossby wave propagation in a barotropic atmosphere. *Dyn. Atmos. Oceans*, **7**, 111–125, https://doi.org/10.1016/0377-0265(83)90013-1.

Lin, H., and G. Brunet, 2009: The influence of the Madden–Julian Oscillation on Canadian wintertime surface air temperature. *Mon. Wea. Rev.*, **137**, 2250–2262, https://doi.org/10.1175/2009MWR2831.1.

——, and ——, 2018: Extratropical response to the MJO: Nonlinearity and sensitivity to the initial state. *Mon. Wea. Rev.*, **75**, 219–234, https://doi.org/10.1175/JAS-D-17-0189.1.

——, ——, and J. Derome, 2009: An observed connection between the North Atlantic Oscillation and the Madden–Julian Oscillation. *J. Climate*, **22**, 364–380, https://doi.org/10.1175/2008JCLI2515.1.

——, ——, and R. Mo, 2010: Impact of the Madden–Julian Oscillation on wintertime precipitation in Canada. *Mon. Wea. Rev.*, **138**, 3822–3839, https://doi.org/10.1175/2010MWR3363.1.

Mo, R., and D. Straus, 2002: Statistical–dynamical seasonal prediction based on principal component regression of GCM ensemble integrations. *Mon. Wea. Rev.*, **130**, 2167–2187, https://doi.org/10.1175/1520-0493(2002)130<2167:SDSPBO>2.0.CO;2.

Murphy, A., 1969: On the ranked probability skill score. *J. Appl. Meteor.*, **8**, 988–989, https://doi.org/10.1175/1520-0450(1969)008<0988:OTPS>2.0.CO;2.

——, 1971: A note on the ranked probability skill score. *J. Appl. Meteor.*, **10**, 155–156, https://doi.org/10.1175/1520-0450(1971)010<0155:ANOTRP>2.0.CO;2.

Pegion, K., and Coauthors, 2019: The Subseasonal Experiment (SubX): A multi-model subseasonal prediction experiment. *Bull. Amer. Meteor. Soc.*, https://doi.org/10.1175/BAMS-D-18-0270.1, in press.

Rodney, M., H. Lin, and J. Derome, 2013: Data analysis and representation on a general domain using eigenfunctions of Laplacian. *Mon. Wea. Rev.*, **141**, 2897–2909, https://doi.org/10.1175/MWR-D-12-00221.1.

Ropelewski, C., and M. Halpert, 1986: North American precipitation and temperature patterns associated with El Niño/Southern Oscillation (ENSO). *Mon. Wea. Rev.*, **114**, 2352–2362, https://doi.org/10.1175/1520-0493(1986)114<2352:NAPATP>2.0.CO;2.

Rukhovets, L. V., H. van den Dool, and A. Barnston, 1998: Forecast-observation pattern relationships in NCEP medium range forecasts of non-winter Northern Hemisphere 500-mb height fields. *Atmos.–Ocean*, **36**, 55–70, https://doi.org/10.1080/07055900.1998.9649606.

Saito, N., 2008: Subseasonal prediction of wintertime North American surface air temperature during strong MJO events. *J. Climate*, **28**, 7420–7436, https://doi.org/10.1175/MWR-D-12-00221.1.

Smith, C., and P. Sardeshmukh, 2000: The effect of ENSO on intraseasonal variance of surface temperatures in winter. *Int. J. Climatol.*, **20**, 1543–1557, https://doi.org/10.1002/1097-0088(20001115)20:13<1543::AID-JOC579>3.0.CO;2-A.

Smith, T., and R. Livezey, 1999: GM systematic error correction and specification of the seasonal mean Pacific–North America region atmosphere from global SSTs. *J. Climate*, **12**, 273–288, https://doi.org/10.1175/1520-0442-12.1.273.

Tippett, M., M. Barlow, and B. Lyon, 2003: Statistical correction of central southwest Asia winter precipitation simulations. *Int. J. Climatol.*, **23**, 1421–1433, https://doi.org/10.1002/joc.947.

——, A. Barnston, and A. Robertson, 2007: Estimation of seasonal precipitation tercile-based categorical probabilities from ensembles. *J. Climate*, **20**, 2210–2228, https://doi.org/10.1175/JCLI4108.1.

——, ——, and T. DelSole, 2010: Comments on "Finite samples and uncertainty estimates for skill measures for seasonal prediction." *Mon. Wea. Rev.*, **138**, 1487–1493, https://doi.org/10.1175/2009MWR3214.1.

——, T. DelSole, and A. G. Barnston, 2014: Reliability of regression-corrected climate forecasts. *J. Climate*, **27**, 3393–3404, https://doi.org/10.1175/JCLI-D-13-00565.1.

Vigaud, N., A. Robertson, and M. Tippett, 2017a: Multimodel ensembling of subseasonal precipitation forecasts over North America. *Mon. Wea. Rev.*, **145**, 3913–3928, https://doi.org/10.1175/MWR-D-17-0092.1.

——, ——, ——, and N. Acharya, 2017b: Subseasonal predictability of boreal summer monsoon rainfall from ensemble forecasts. *Front. Environ. Sci.*, **5**, https://doi.org/10.3389/fenvs.2017.00067.

——, M. Tippett, and A. Robertson, 2018: Probabilistic skill of subseasonal precipitation forecasts for the East Africa–West Asia sector during September to May. *Mon. Wea. Rev.*, **146**, 2559–2577, https://doi.org/10.1175/MWR-D-18-0058.1.

Vitart, F., and Coauthors, 2017: The Subseasonal to Seasonal (S2S) prediction project database. *Bull. Amer. Meteor. Soc.*, **98**, 163–173, https://doi.org/10.1175/BAMS-D-16-0017.1.

Ward, N., and A. Navarra, 1997: Pattern analysis of SST-forced variability in ensemble GCM simulations: Examples over Europe and the tropical Pacific. *J. Climate*, **11**, 711–743, https://doi.org/10.1175/1520-0442(1997)010<2210:PAOSFV>2.0.CO;2.

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

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

Wilks, D., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction*. International Geophysics Series, Vol. 59, Elsevier, 467 pp.

——, 2009: Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteor. Appl.*, **16**, 361–368, https://doi.org/10.1002/met.134.

——, and T. Hamill, 2007: Comparison of ensemble MOS methods using GFS reforecasts. *Mon. Wea. Rev.*, **135**, 2379–2390, https://doi.org/10.1175/MWR3402.1.

Yao, W., H. Lin, and J. Derome, 2011: Submonthly forecasting of winter surface air temperature in North America based on organized tropical convection. *Atmos.–Ocean*, **49**, 51–60, https://doi.org/10.1080/07055900.2011.556882.