Spatial Correction of Multimodel Ensemble Subseasonal Precipitation Forecasts over North America Using Local Laplacian Eigenfunctions

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 1 May 2019, in final form 7 October 2019)

ABSTRACT

The extent to which submonthly forecast skill can be increased by spatial pattern correction is examined in probabilistic rainfall forecasts of weekly and week-3–4 averages, constructed with extended logistic regression (ELR) applied to three ensemble prediction systems from the Subseasonal-to-Seasonal (S2S) project database. The new spatial correction method projects the ensemble-mean rainfall neighboring each grid point onto Laplacian eigenfunctions and then uses those amplitudes as predictors in the ELR. Over North America, individual and multimodel ensemble (MME) forecasts that are based on spatially averaged rainfall (e.g., first Laplacian eigenfunction) are characterized by good reliability, better sharpness, and higher skill than those using the gridpoint ensemble mean. The skill gain is greater for week-3–4 averages than week-3 leads and is largest for MME week-3–4 outlooks that are almost 2 times as skillful as MME week-3 forecasts over land. Skill decreases when using more Laplacian eigenfunctions as predictors, likely because of the difficulty in fitting additional parameters from the relatively short common reforecast period. Higher skill when increasing reforecast length indicates potential for further improvements. However, the current design of most subseasonal forecast experiments may prove to be a limit on the complexity of correction methods. Relatively high skill for week-3–4 outlooks with winter starts during El Niño and MJO phases 2–3 and 6–7 reflects particular opportunities for skillful predictions.

1. Introduction

Subseasonal-to-seasonal forecasting (lead times between 2 weeks and 2 months) is currently the focus of intense research efforts within the World Weather Research Programme–World Climate Research Programme (WWRP/WCRP) Subseasonal-to-Seasonal (S2S) prediction project (Vitart 2014), the aims of which include developing well-calibrated probabilistic subseasonal forecasts. One of the challenges of assessing probabilistic skill of S2S forecasts is that reforecast ensembles generally contain fewer ensemble members than in the seasonal forecasting case, so a straightforward computing of probabilities by counting of reforecast ensemble members exceeding a chosen threshold leads to large errors. For instance, NCEP and CMA reforecast archives from the S2S database used in this study have only four members, thus the reforecast probabilities obtained by counting can only take the values of 0%, 25%, 50%, 75%, and 100%, which are coarse estimates. By contrast, distributional regression is well suited to probability forecasting and regression models are more skillful than straight counting for small ensemble size in the seasonal forecasting context (Tippett et al. 2007). Extended logistic regression (ELR), which ensures consistent forecast probabilities across a range of thresholds (Wilks 2009), was thus chosen in Vigaud et al. (2017a) to design a multimodel ensemble (MME) prediction system for submonthly forecasts from three ensemble prediction systems, or EPSs [European Centre for Medium-Range Weather Forecasts (ECMWF); National Centers for Environmental Prediction (NCEP) Climate Forecast

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

DOI: 10.1175/MWR-D-19-0134.1

System, version 2 (CFSv2); and China Meteorological Administration (CMA)] in the S2S database (Vitart et al. 2017). As in seasonal (3-month averages) and medium range (out to 15 days) forecasting (Robertson et al. 2004; Hamill 2012), these ELR-based forecasts show enhanced probabilistic forecast skill through calibration and multimodel ensembling across S2S time scales in different regions including North America (Vigaud et al. 2017a,b, 2018). In the ELR method, 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). In addition, local regression relationships are prone to sampling uncertainties that can further lead to spatially noisy forecasts, hence there might be potential for improvements by including spatial information. This study thus examines the extent to which probabilistic skill can be improved by spatial correction relative to those of baseline ELR forecasts in Vigaud et al. (2017a) for week-1 through week-4 (i.e., $[d + 1; d + 7]$ through $[d + 22; d + 28]$ targets for a forecast initialized on day d) and week-3–4 (i.e., $[d + 15; d + 28]$ targets) precipitation tercile forecasts over continental North America.

Multiple linear regressions such as principal component regressions (Mo and Straus 2002) or canonical correlation analysis (Barnston and Ropelewski 1992), are well suited for model output statistics and correct systematic errors in pattern positions and amplitudes of 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, the Gaussian assumption made by these methods still needs to be tested at subseasonal time scales. Among other approaches, Laplacian eigenfunction decomposition, which has been recently applied to climate analysis (Saito 2008; DelSole and Tippett 2015), makes no assumption on the data and represents an attractive alternative to summarize spatial information by filtering out small-scale variability. Depending only on the geometry of the domain, Laplacian eigenfunctions are well suited for multimodel studies because they are uniformly defined across models (DelSole and Tippett 2015). This study thus examines the extent to which an existing ELR-based probabilistic prediction system for submonthly rainfall forecasts (Vigaud et al. 2017a) can be improved by enabling spatial pattern correction through the decomposition of ensemble mean rainfall neighboring each grid point using locally defined Laplacian eigenfunctions. Similarly to the existing ELR, the Laplacian-ELR (L-ELR) approach is applied to individual model forecasts then averaged with equal weights to produce MME precipitation tercile probability forecasts for weekly and week-3–4 averages.

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

Attributes	ECMWF	NCEP	CMA
Time range	Days 0–46	Days 0–44	Days 0–60
Resolution	Tco639/319 L91	T126L64	T106L40
Ensemble size	51	16	4
Frequency	2 per week	daily	daily
Reforecasts (RFC)	On the fly	Fixed	Fixed
RFC length	Past 20 yr	1999–2010	1994–2014
RFC frequency	2 per week	Daily	Daily
RFC size	11	4	4

The methods and data are presented in section 2. The skill of L-ELR forecasts initialized during January–March (JFM; winter) and July–September (JAS; summer) is next investigated over North America and compared with those obtained from the existing ELR model in section 3, alongside skill relationships to ENSO conditions and Madden–Julian oscillation (MJO) phases. A summary and conclusions are gathered in section 4.

2. Data and methods

a. Observation and model datasets

Observation and model datasets are the same as in Vigaud et al. (2017a), which the following data description parallels in the next two paragraphs. Week-1 $[d + 1; d + 7]$, week-2 $[d + 8; d + 14]$, week-3 $[d + 15; d + 21]$, week-4 $[d + 22; d + 28]$, and week-3–4 $[d + 15; d + 28]$ targets for a forecast issued on day d were computed from daily rainfall from the ECMWF, NCEP, and CMA hindcasts (referred to as reforecasts in the following) acquired from the S2S database (Vitart et al. 2017). These EPSs have distinct resolutions, numbers of ensemble members, and reforecasts lengths as indicated in Table 1, but in the S2S database they are all archived on the same 1.5° grid. ECMWF is the only model with reforecasts (11 members) generated two times per week (Mondays and Thursdays) on the fly (i.e., new reforecast sets are generated twice-weekly, with the latest model version used to produce real-time ensemble forecasts for the following 46 days). By contrast, NCEP and CMA reforecasts are issued four times daily using the same fixed version of their respective models. Such differences are inherent to the two configurations used by the different centers producing reforecasts archived in the S2S database. Weekly accumulated precipitation from ECMWF reforecasts generated for Thursday starts in 2016 is used in the following analysis, contrasting with Vigaud et al. (2017a) based on Monday starts. Similarly,

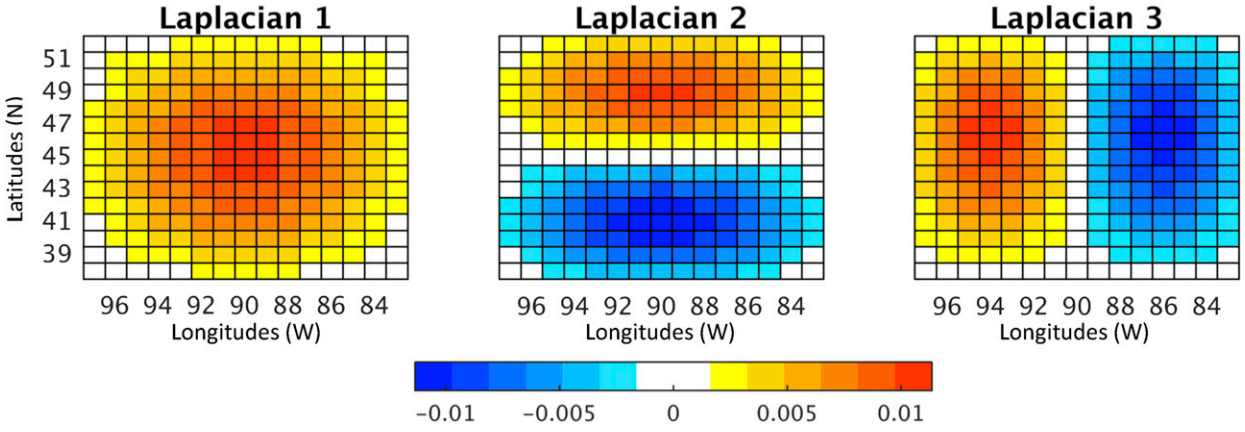

FIG. 1. First three Laplacians at 45°N, 90°W computed on a geographical box of 15 neighboring grid points in latitude and longitude.

however, ECMWF Thursday starts in 2016 also comprise different model cycles (CY41R1, CY41R2 and CY43R1) across the whole calendar year. Summer starts (July–September) are based on model version CY41R2, whereas the winter starts (January–March) are based on model cycles CY41R1 and CY41R2. The main change between CY41R1 and CY41R2 (introduced on 8 March 2016) was a doubling of the spatial atmospheric resolution. This change may have improved the skill during March, but its impact on the seasonally averaged skill differences between ELR and L-ELR methods is expected to be minor. NCEP and CMA four-member daily reforecasts are then sampled for ECMWF 2016 Thursday calendar start dates across each year, thus allowing to design a multimodel ensemble based on exactly the same issuance dates across models, similarly to the probabilistic skill analysis of precipitation forecast from Vigaud et al. (2017a), based on the same three-models subset. The reforecasts from all three EPSs are available from 1999 to 2010, which is the period used in our study. There are thus 144 reforecasts for the JFM and JAS seasons (12 starts over 12 years) and each model. ECMWF reforecasts over the 1997–2014 period are also used to test the effect of sample variability on forecast skill by increasing reforecast length (section 3b). To produce comparable sets of precipitation tercile category probabilities (referred to as forecasts in the following), S2S model data were all interpolated spatially onto Global Precipitation Climatology Project (GPCP) 1° horizontal grid. Forecast probabilities are computed for the three models individually and then averaged to form MME precipitation tercile category probabilities (referred to as MME forecasts) the skill of which is assessed over continental North America (i.e., land points between 20° and 50°N) for winter (JFM) and summer (JAS) starts.

Daily estimates from the GPCP (Huffman et al. 2001; Huffman and Bolvin 2012), version 1.2, available on a 1° grid from 1996 to 2015 are the observational data used to calibrate and verify the reforecasts over 1999–2010.

b. *Local Laplacian eigenfunction decomposition*

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

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

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

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

where

$$dx_i \equiv \lambda_i - \lambda_{i-1} \quad \text{and} \quad dy_i \equiv \frac{2(\phi_i - \phi_{i-1})}{\cos\phi_i + \cos\phi_{i-1}}. \tag{3}$$

For each grid point of the North American domain, the matrix representation of Eq. (2) with Dirichlet boundary conditions is formed for the 15° × 15° box centered on that grid point. The eigenvectors of this 225 × 225 matrix are then computed. For each model, grid point, start and lead, reforecasts are next projected onto the first three Laplacian eigenfunctions shown in Fig. 1 to be used as predictors in the ELR model. Because eigenfunctions are only unique up to a multiplicative constant, the projection is done with area weighting as

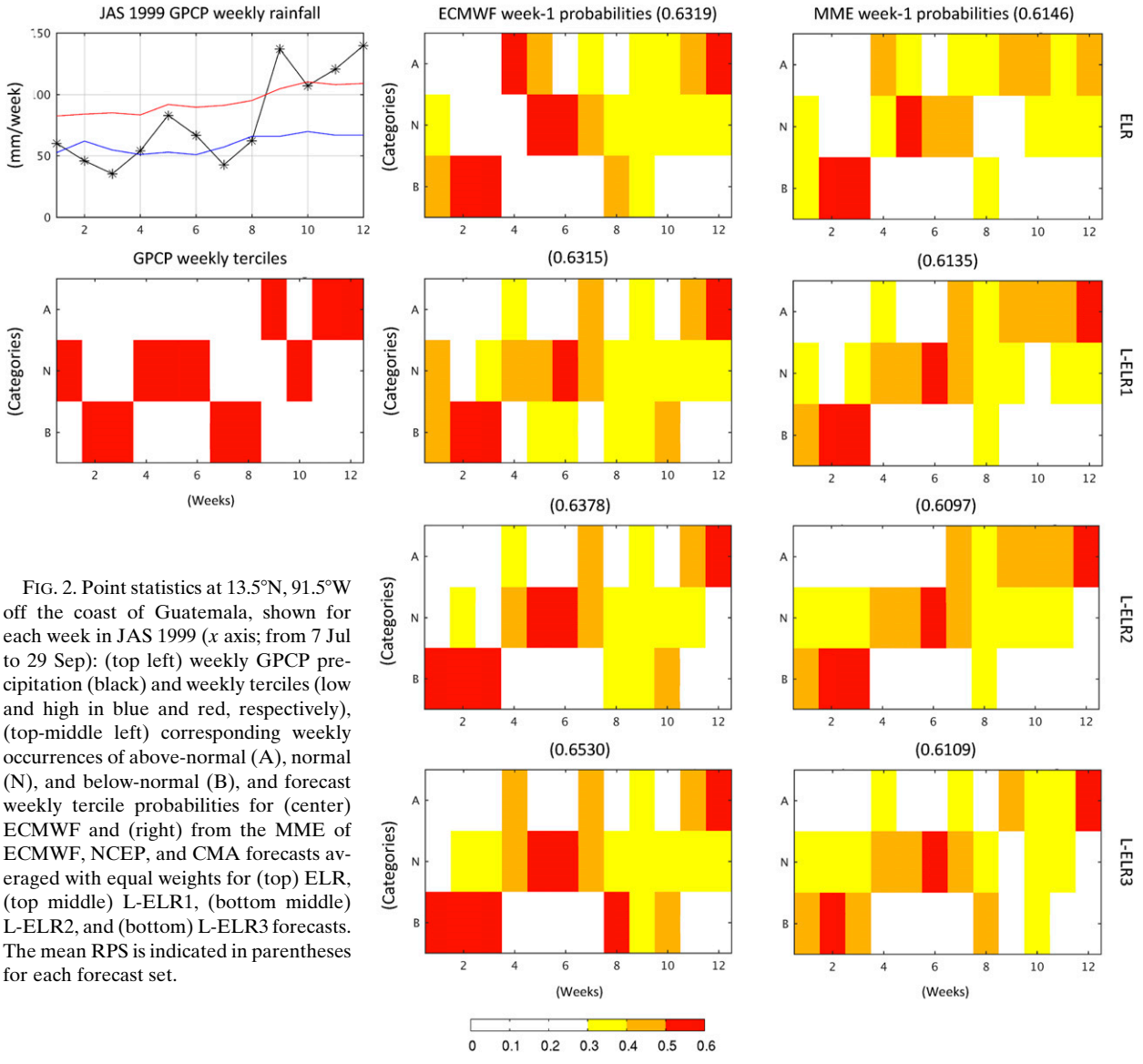

FIG. 2. Point statistics at 13.5°N, 91.5°W off the coast of Guatemala, shown for each week in JAS 1999 (x axis; from 7 Jul to 29 Sep): (top left) weekly GPCP precipitation (black) and weekly terciles (low and high in blue and red, respectively), (top-middle left) corresponding weekly occurrences of above-normal (A), normal (N), and below-normal (B), and forecast weekly tercile probabilities for (center) ECMWF and (right) from the MME of ECMWF, NCEP, and CMA forecasts averaged with equal weights for (top) ELR, (top middle) L-ELR1, (bottom middle) L-ELR2, and (bottom) L-ELR3 forecasts. The mean RPS is indicated in parentheses for each forecast set.

in DelSole and Tippett (2015) such that the chosen constant satisfies that the area average of the squared eigenfunction is equal to one. As shown by values plotted in Fig. 1 that reflect corresponding geographical weights given to reforecasts grid points when projected on Laplacians, the first Laplacian eigenfunction represents a weighted spatial average, while the second and third correspond to meridional and zonal gradients, respectively. The local Laplacian eigenvectors used here differ from those in DelSole and Tippett (2015) in that the ones here are computed on rectangular domains and satisfy an explicit Dirichlet boundary condition. The Laplacian eigenvectors in DelSole and Tippett (2015) can be computed on arbitrary domains and satisfy a nonlocal boundary condition (Saito 2008).

c. *Extended logistic regression model*

The method 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:

$$\ln\left(\frac{p}{1-p}\right) = f(\overline{x_{\text{ens}}}) + g(q), \tag{4}$$

with $f = b_0 + b_1\overline{x_{\text{ens}}}$ and $g = b_2 q$, where b_0 and b_1 are regression coefficients and $\overline{x_{\text{ens}}}$ is the gridpoint ensemble mean precipitation. Cumulative probabilities obtained

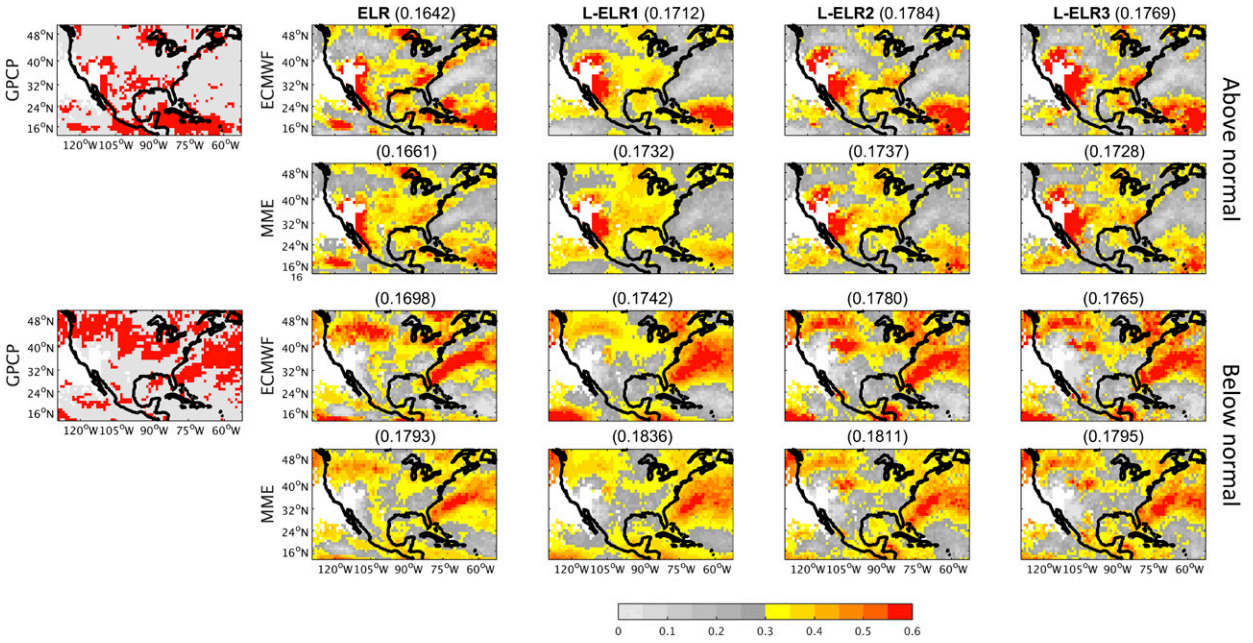

FIG. 3. Observed GPCP (top left) above- and (bottom-middle left) below-normal precipitation tercile probabilities for 7 Jul 1999 start, together with those forecast by ECMWF and the multimodel ensemble (MME) of ECMWF, CFS, and CMA models from ELR and L-ELR1–3 forecasts, as indicated by the labels. White shadings correspond to the dry-mask equivalent, and mean Brier score averages over the entire domain are indicated in parentheses for each forecast.

from Eq. (4) for smaller predictand thresholds cannot exceed those for larger thresholds (Vigaud et al. 2017a) yielding logically consistent sets of forecasts (Wilks 2009). Precipitation tercile category probabilities (ELR forecasts) are here computed using ELR for the 33rd and 67th precipitation percentiles.

Observed climatological weekly tercile categories derived from GPCP weekly cumulated precipitation estimates are defined based on 3-week windows including the target week and one week on either side, separately at each grid point for each start in JFM (7 January–31 March Thursday start dates) and JAS (7 July–29 September Thursday start dates), and each lead (weeks 1–4) following a leave-one-year-out approach (i.e., using 33 weeks from 11 years). ELR forecasts are produced only where and when the lower tercile (33rd percentile) is nonzero (i.e., ''dry mask'' equivalent). In such dry areas where more than one-third of the observations are zero, other categorical forecast targets such as rain/no rain, or above or below the median may be more suitable. Observed climatological biweekly terciles are defined on 6-week windows centered on week-3–4 targets (i.e., $[d + 15; d + 28]$).

For each S2S model, the same pool of weeks on which terciles are defined under cross validation (i.e., 3-week windows centered on the target week, over 11 years) are used to train the ELR model out-of-sample by fitting forecasts equations at each grid point, lead, and calendar start date separately. The regression coefficients thus obtained are then used to predict terciles probabilities for the left-out year (validation set) that are averaged across models with equal weights to produce an MME of the individual forecast probabilities (MME forecasts); see Vigaud et al. (2017a) for more details.

To correct forecasts spatially, the Laplacian eigenfunction decomposition of neighboring ensemble mean precipitation (Lap) is used in Eq. (4) instead of the gridpoint average $\overline{x_{\text{ens}}}$:

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

where $f = b_0 + \sum_{i=1}^{n} b_i \times \text{Lap}_i$ and $g = b_{n+1}q$, with Lap_i corresponding to the projection of the ensemble mean precipitation of 15 grid points neighboring each location on the ith Laplacian eigenvector defined on this geographical box. ELR models based on n eigenvectors to produce tercile probabilities will be referred to as L-ELRn forecasts for n = 1–3: L-ELR2 using the spatial average and meridional gradient provided by the first two eigenvectors, for instance.

Similar to Fig. 2 in Vigaud et al. (2017a), but for 2016 ECMWF Thursday starts, Fig. 2 shows GPCP

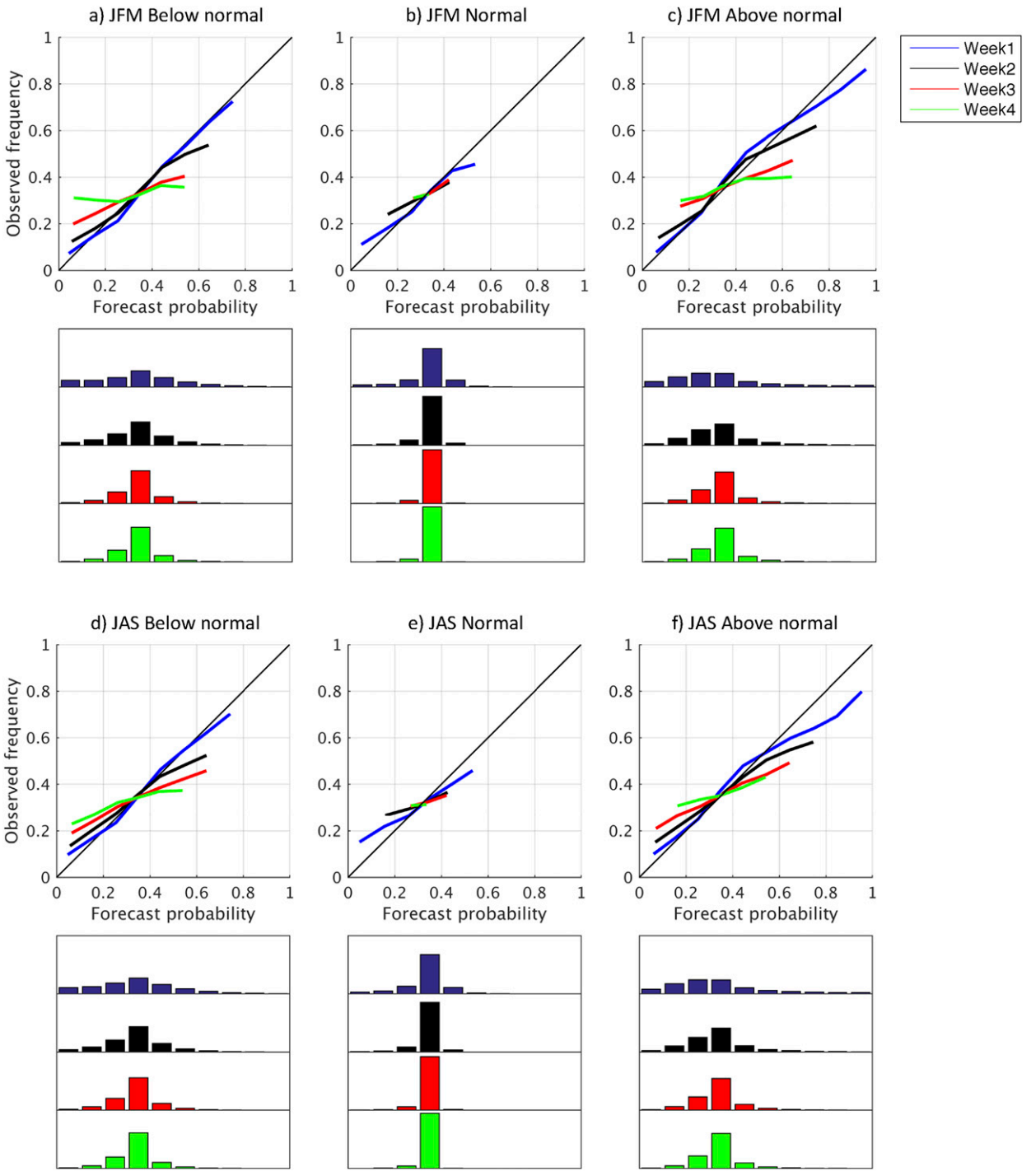

FIG. 4. Reliability diagrams for all three categories [(left) below normal, (center) normal, and (right) above normal] from ECMWF L-ELR1 forecasts with starts in (a)–(c) JFM and (d)–(f) JAS from week-1 to week-4 leads (colors). Forecast frequencies of issuance are shown as bins in histograms under the 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.

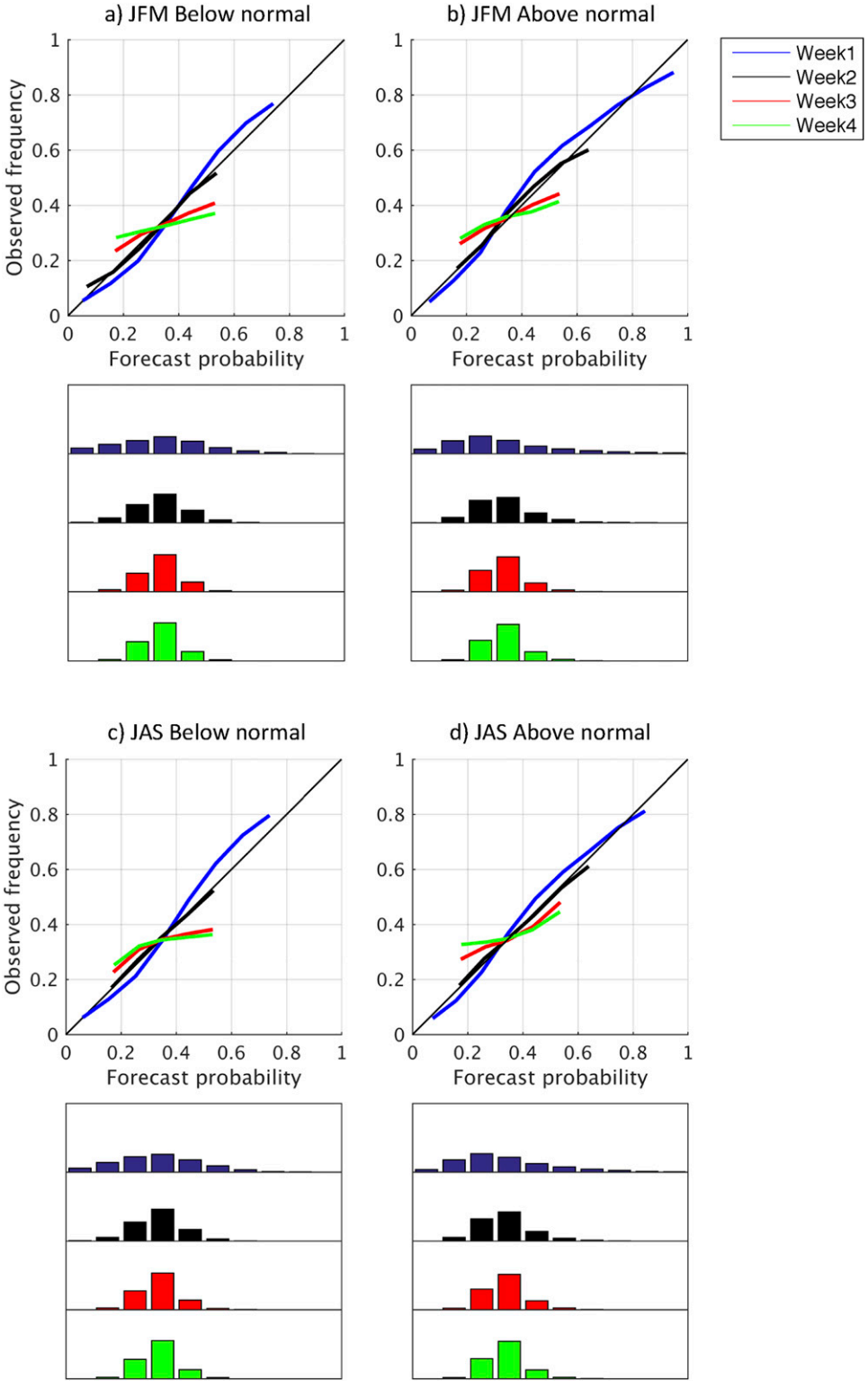

FIG. 5. As in Fig. 4, but for the (left) below-normal and (right) above-normal categories from the MME of ECMWF, NCEP, and CMA L-ELR1 forecasts with (a),(b) JFM and (c),(d) JAS starts.

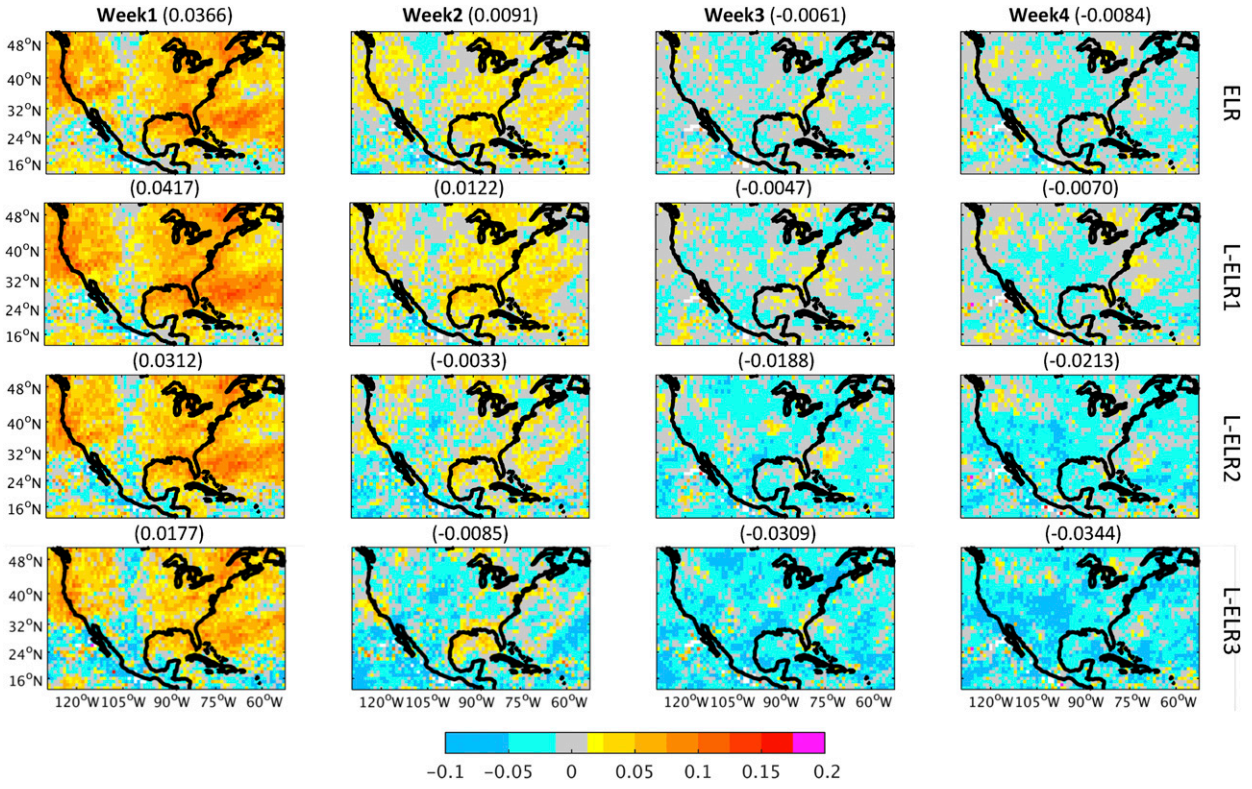

FIG. 6. RPSS for ECMWF tercile precipitation from (top) ELR and (top middle) L-ELR1, (bottom middle) L-ELR2, and (bottom) L-ELR3 forecasts for JFM starts and different leads of (left) 1, (left center) 2, (right center) 3, and (right) 4 weeks. Mean RPSS is indicated in parentheses for each forecast.

observations with ELR and L-ELR forecast probabilities at week-1 lead for all starts in JAS 1999 at a grid point off the Pacific coast of Guatemala (13.5°N, 91.5°W). ECMWF category forecasts display highest weekly probabilities mostly consistent with observed tercile categories, and are more skillful than those from NCEP and CMA (not shown), as well as the three models averaged with equal weights, directly reflecting the lesser performances of NCEP and CMA relative to ECMWF. Higher ranked probability score (RPS) values for L-ELR than ELR reflect modest skill improvements for ECMWF but not for the MME; however, RPS differences are too low to be significant.

Probability maps from ELR and L-ELR1–3 forecasts for 7 July 1999 start (Fig. 3) display highest probabilities geographically consistent with GPCP, MME forecasts being spatially smoother than ECMWF with broader areas of lower maximum probabilities. Overall, these reflect wetter-than-normal conditions in the tropics, where convective rainfall is typical of the wet season in the Intra-American Seas (IAS) and American monsoon regions, whereas below-normal probabilities in the midlatitudes are consistent with dry conditions during summer over these regions of North America. L-ELR1 forecasts have less-noisy probabilities relative to the baseline ELR, similarly to L-ELR2–3. Brier scores (Brier 1950), which are between 0 for a perfect forecast and 1 for no forecast skill, are used to roughly verify these probability forecasts over the whole domain. Lower Brier scores for ELR than L-ELR forecasts for ECMWF and the MME further illustrate the added value of spatial correction relative to calibration without knowledge of neighboring grid points.

d. Skill metrics and significance testing

Reliability diagrams are first computed for all grid points over continental North America between 20° and 50°N in latitudes to evaluate the reliability, but also resolution as well as sharpness (Wilks 1995; Hamill 1997), of ELR and L-ELR tercile category precipitation forecasts. To complement reliability diagrams with spatial information, maps of ranked probability skill scores (RPSS; Epstein 1969; Murphy 1969, 1971; Weigel et al. 2007) are next used to quantify to which extent calibrated predictions are improved in comparison to climatological frequencies. Generalized skill scores tend to be not strictly proper (Gneiting and Raftery 2007); however, RPSS remains one of the most commonly used

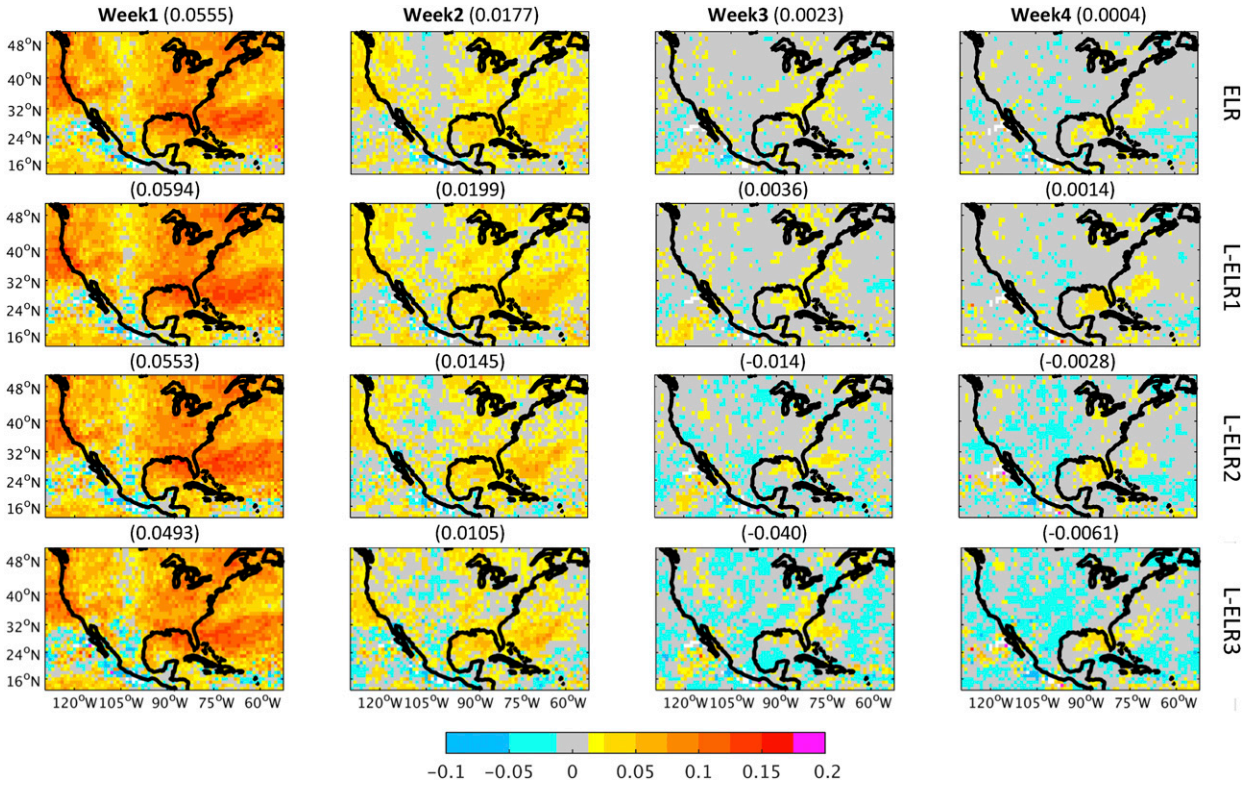

FIG. 7. As in Fig. 6, but for the MME of ECMWF, CFS, and CMA forecasts.

skill scores, which is generally preferred to others that are also sensitive to distance (Daan 1985; Wilks 1995; Weigel et al. 2007), and its values tend to be small. For joint-Gaussian forecasts and observations, a reliable deterministic forecast with correlation r will have an RPSS of approximately $1 - (1 - r^2)^{1/2}$, meaning that an RPSS value of 0.1 corresponds to a correlation of about 0.44 (Tippett et al. 2010).

The statistical significance of area averages of RPSS during specific ENSO and MJO phases is assessed by a permutation test. Area averages of RPSS that exceed the 90th percentile from 100 000 permutations of forecasts with JFM starts are statistically significant at the 0.1 significance level.

3. Results

a. Weekly averages

Reliability diagrams for weekly ECMWF L-ELR1 forecasts with starts in the JFM and JAS seasons (Fig. 4) are very similar to those from ELR for 2016 ECMWF Monday starts in Vigaud et al. (2017a) with good reliability and resolution for week 1, as shown by blue lines near the diagonal and away from the 0.33 horizontal line (not plotted), respectively. Histograms spread across all bins for week-1 forecasts characterize high sharpness, except for the normal category. This is consistent with seasonal forecasts of the likelihood of the near-normal category, which cannot be greatly sharpened beyond the climatology forecast (Kharin and Zwiers 2003) and are thus not much more skillful than the climatology (van den Dool and Toth 1991). The distribution of forecast frequencies are skewed toward climatology (0.33, i.e., fourth bin) with increasing leads, consistently with decreasing slopes from week 2, when reliability and resolution drop with higher skill in winter than summer and little skill visible at higher leads. NCEP and CMA display similar results but are less skillful.

Greater slopes for the MME (Fig. 5) reflect better reliability and resolution. Similarly to ELR in Vigaud et al. (2017a), reliability and sharpness are degraded by multimodel ensembling for L-ELR1 forecasts, with MME histograms slightly less spread than for the ECMWF from week 2, reflecting the lesser performances of NCEP and CMA compared to ECMWF.

As for ELR forecasts in Vigaud et al. (2017a), the northwestern and eastern United States exhibit maximum positive RPSS values over land for all week-1 ECMWF forecasts with starts in JFM (Fig. 6), where skill

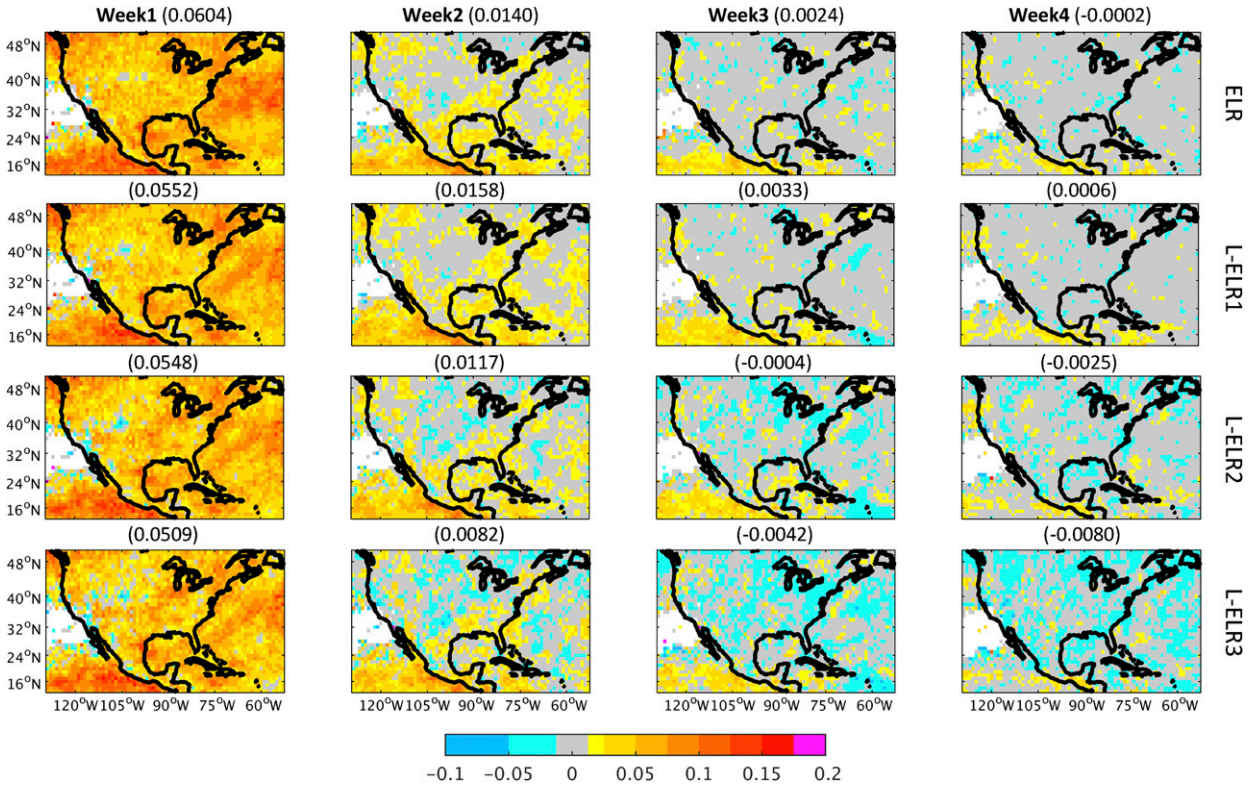

FIG. 8. As in Fig. 7, but for starts during the JAS season. White shadings correspond to the dry-mask equivalent.

persists in week 2 but with less magnitude. These maximums might be associated with the influence of midlatitude depressions affecting rainfall in both regions during winter (Barnston and Livezey 1987). Near-zero or negative values are found everywhere at higher leads. RPSS decreases from L-ELR1 to L-ELR3, but higher mean RPSS for L-ELR1 than ELR indicates more skill.

Relative to the best model (ECMWF), RPSS is increased everywhere by multimodel combination (Fig. 7), which damps negative skill values at all leads. The Southwest and the eastern United States still exhibit maximum positive RPSS in week 2 over land, where positive skill only remains over the U.S. East Coast and northeastern regions of the Gulf of Mexico in week 3 and could be explained by local rainfall relationships to jet-stream modulations and shifts of storm tracks in winter (Barnston and Livezey 1987; Monteverdi and Null 1998). MME RPSS levels decrease from L-ELR1 to L-ELR3 but increase from ELR to L-ELR1 with greater gain than ECMWF at all leads. Skill is lower in summer (Fig. 8), with maximum RPSS values north and south of the subcontinent, south of 24°N and north of 40°N in the Pacific, over the IAS and U.S. East Coast in week 1. These skill patterns are consistent with the occurrences of convective rainfall and storms typical of the wet season within the IAS and American monsoon regions, as well as prevailing dry conditions over North America midlatitudes in summer. Low skill at higher leads do not allow to identify any increase from ELR to L-ELR1 and RPSS decreases with more Laplacians, reflecting the low predictability of tropical rainfall.

Figure 9 shows spatial averages over North America between 20° and 50°N of the percentages of forecasts different from climatology in Figs. 4 and 5, which is an indication of sharpness, together with spatial averages of RPSS for ECMWF and MME forecasts with starts in JFM. Sharpness and RPSS decrease with lead and reflect low skill after week 2, when mean RPSS is only positive for ELR and L-ELR1 MME forecasts. L-ELR1 forecasts have higher sharpness and RPSS than those from ELR for both ECMWF and the MME at all leads, confirming higher skill when using spatially averaged rainfall (i.e., Laplacian 1) than the gridpoint mean as predictor. Increasing the number of predictors in L-ELR2–3 degrades the forecasts at all leads in terms of RPSS while sharpness is increased from ELR and L-ELR1 even at week-1 lead and, together with decreasing skill in Figs. 6–8, suggests overconfidence and reduced reliability. This overconfidence can be related to the sensitivity of regression methods to sample

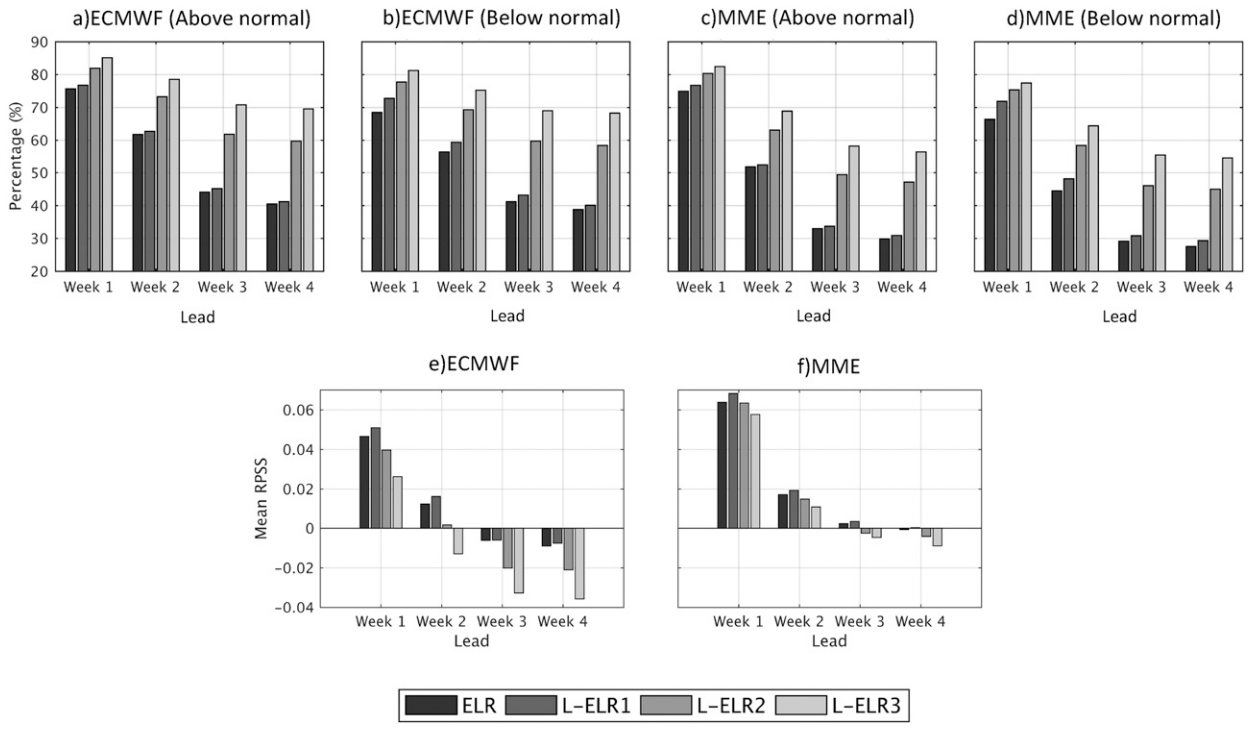

FIG. 9. Percentages of forecasts outside the fourth bin (0.33) in Figs. 4 and 5 for week-1–4 forecasts from (a),(b) ECMWF and (c),(d) the MME for the below- and above-normal categories as labeled, along with (e),(f) mean RPSS averaged over continental North America between 20° and 50°N latitudes, for ELR and L-ELR1–3 precipitation tercile forecasts with JFM starts.

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 here for training at each start date based on the same pool of weeks used to define terciles under cross-validation (three reforecasts over 11 years) to produce weekly forecasts does not allow to significantly satisfy the rule of thumb of having approximately ten samples per explanatory variables, beyond two predictors (i.e., 1 Laplacian). This aspect is further investigated for week-3–4 outlooks in the following.

b. Week-3–4 outlooks

Reliability diagrams for 2-week week-3–4 L-ELR1 outlooks with JFM starts (Fig. 10) are comparable to those of ELR in Vigaud et al. (2017a), with low sharpness but greater slopes than weekly forecasts (Figs. 4 and 5), thus better reliability and resolution. Greater slopes for week-3–4 forecasts than weekly forecasts indicate increased gain from multimodel ensembling.

Week-3–4 MME outlooks have higher RPSS values (Fig. 11) than week-3 and week-4 forecasts over the northeastern, western, and southwestern United States, across the IAS and Florida in JFM, and over the West United States and east Pacific for JAS starts. These skill patterns are consistent with winter rainfall relationships to the modulations of the jet stream, storm tracks and atmospheric rivers (Barnston and Livezey 1987; Monteverdi and Null 1998; Dettinger 2011; Ralph et al. 2011; Zhang 2013), and the maximum ENSO-related predictability of tropical convective rainfall in the eastern Pacific and surroundings in summer. Areas of skill are spatially broader for L-ELR1 compared to ELR, but skill decreases from L-ELR2 to L-ELR3 likely reflecting limitations from the small sample size. The effect of sample variability can, however, be reduced by increasing reforecast length (Tippett et al. 2014), as shown in Fig. 13 by increased RPSS for ECMWF week-3–4 forecasts, as well as weekly targets (not shown), and enhanced skill gain for L-ELR1 when verifying and extending the pool of reforecasts to 1997–2014. Less improvement is seen in RPSS, however, when verifying forecasts over the 1999–2010 period (Fig. 13 bottom panels).

Over North America in winter, week-3–4 outlooks exhibit systematically higher RPSS and sharpness than week-3 and week-4 forecasts (Fig. 12) with greater RPSS for the MME compared to ECMWF for all forecasts, while the opposite is true for sharpness at all leads. L-ELR1 week-3–4 outlooks are in average

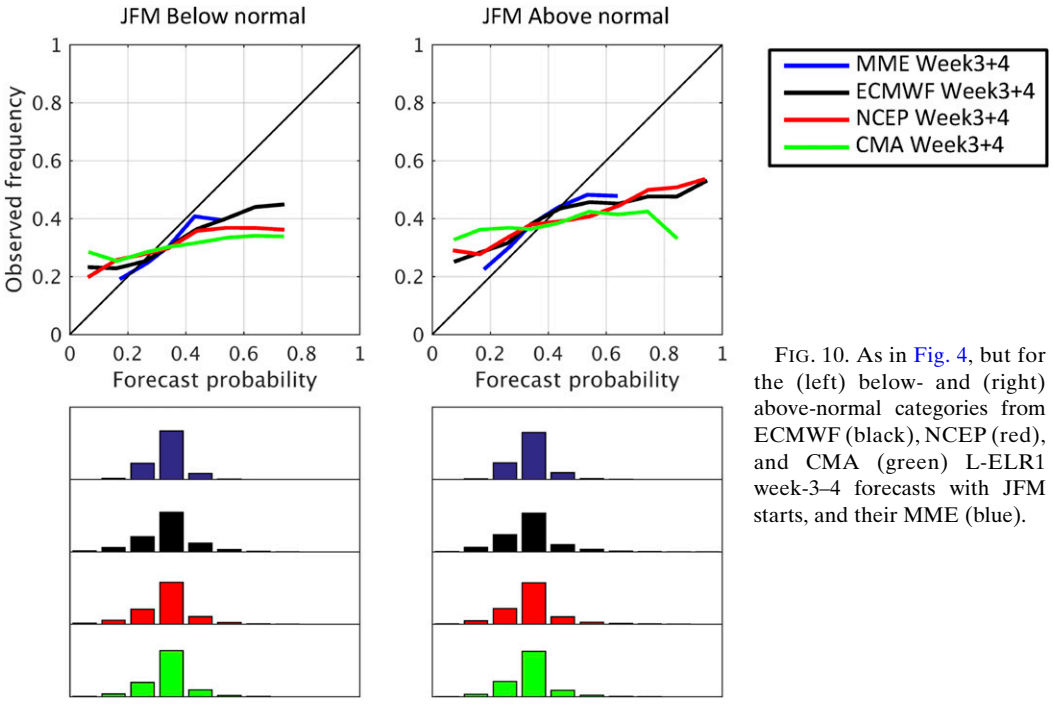

FIG. 10. As in Fig. 4, but for the (left) below- and (right) above-normal categories from ECMWF (black), NCEP (red), and CMA (green) L-ELR1 week-3–4 forecasts with JFM starts, and their MME (blue).

slightly sharper than those from ELR. High levels of sharpness increasing from L-ELR2 to L-ELR3 for both ECMWF and the MME (from 50% to 70%) and lower week3–4 RPSS levels in L-ELR2–3 than ELR and L-ELR1 forecasts might again indicate overconfidence and reflect the small sample size issue. Sharpness levels remain high but are substantially lower when extending ECMWF reforecasts to the 1997–2014 period (not shown), suggesting reduced overconfidence when the sample size is increased. The gain from multimodel ensembling is greater for week3–4 relatively to weekly averages, with larger RPSS differences between ECMWF and the MME in Fig. 12 than in Fig. 9, and maximized for L-ELR1 forecasts displaying twice the RPSS values from week-3 leads.

c. *Skill relationships to ENSO and the MJO*

Figure 14 top panels show week-3–4 MME RPSS values versus probabilities for the below and above

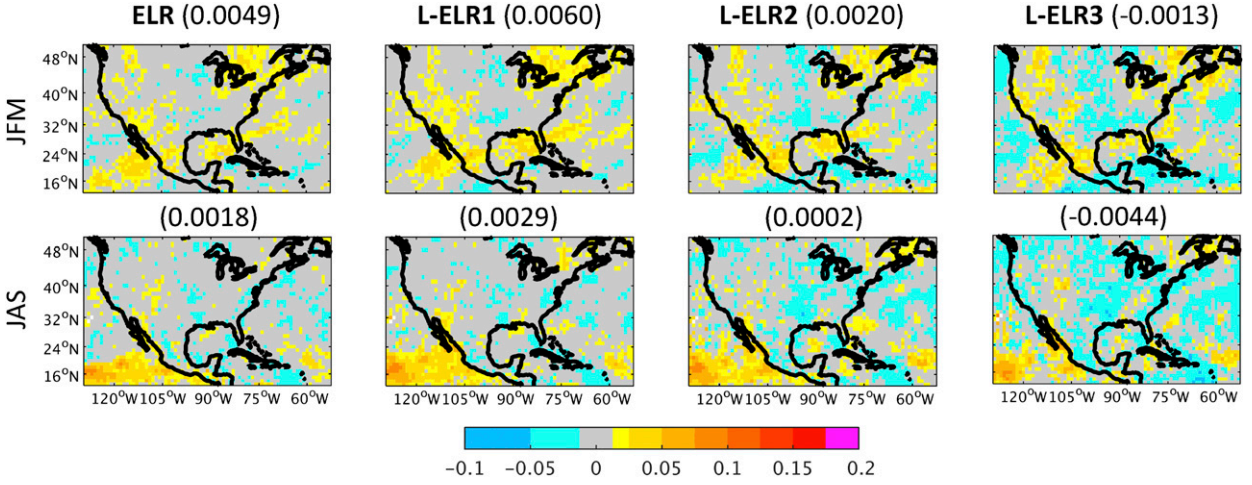

FIG. 11. RPSS for week-3–4 outlooks from the MME of ECMWF, NCEP, and CMA for ELR and L-ELR1–3 tercile precipitation forecasts and all starts during the (top) JFM and (bottom) JAS seasons. Mean RPSS is indicated in parentheses for each forecast.

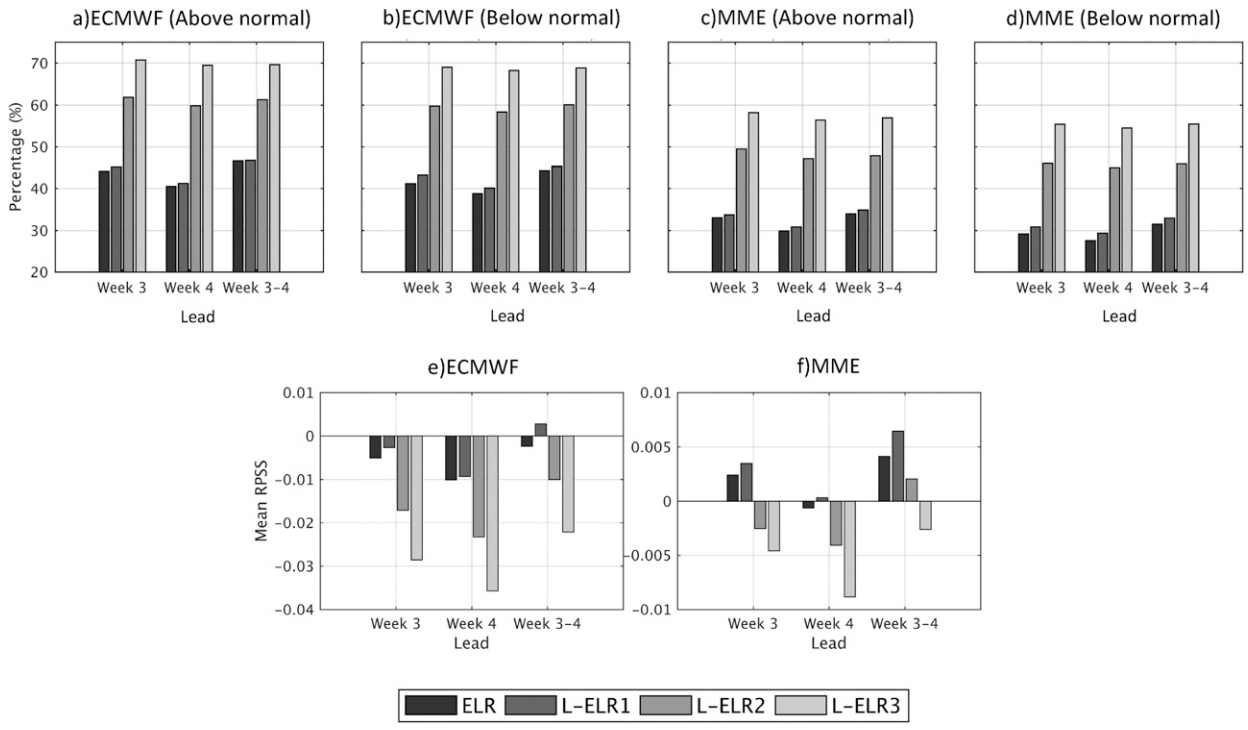

FIG. 12. Percentages of forecasts outside the fourth bin (0.33) in Figs. 4 and 5 for week-3 and week-4 as well as week-3–4 forecasts from (a),(b) ECMWF and (c),(d) the MME for the below- and above-normal categories as labeled, along with (e), (f) mean RPSS averaged over continental North America between 20° and 50°N latitudes, for ELR and L-ELR1–3 precipitation tercile forecasts with JFM starts.

normal categories averaged over land grid points for North America between 20° and 50°N and JFM starts during distinct ENSO conditions (El Niño and La Niña when Niño-3.4 is greater and lower than 0.5, respectively, and neutral conditions otherwise). The highest RPSS values occur during El Niño conditions and correspond to enhanced forecast probabilities for the above normal category. Skill is lower during La Niña, when forecasts tend to indicate drier than normal conditions. This is consistent with maximum RPSS values over the southwestern United States/Mexico in Fig. 11 and increased skill for ELR forecasts over these regions for El Niño (Vigaud et al. 2017a) related to jet-stream modulations and shifts in storm tracks (Barnston and Livezey 1987; Monteverdi and Null 1998). Week-3–4 RPSS increases from ELR to L-ELR1 during El Niño and neutral phases but not during la Niña, with lower RPSS values for L-ELR2–3 in all phases.

Higher mean RPSS and probability ranges across MJO phases (Fig. 14 bottom panels) than for ENSO suggest stronger modulations of skill and probabilistic forecasts. RPSS is highest for forecasts issued during MJO phases 2–3 and 6–7, coinciding with enhanced and reduced forecast probabilities for the above and below normal categories, respectively, except for phase 6, consistent with skill relationships to MJO RMM2 (Vigaud et al. 2017a) maximum during these phases, when the MJO modulates atmospheric rivers and western U.S. rainfall (Zhang 2013). Greater week-3–4 RPSS for L-ELR1 than ELR during most phases again contrasts with lower RPSS values for L-ELR2–3.

4. Discussion and conclusions

The added value of spatial pattern correction on the skill of submonthly precipitation tercile forecasts has been examined for probabilities from extended logistic regression (ELR) when applied to reforecasts from three models (ECMWF, NCEP and CMA) in the S2S database over the common 1999–2010 period. Spatial information is summarized in the ELR model by projecting the ensemble mean precipitation over 15 grid points in latitude and longitude neighboring each location onto the Laplacian eigenfunctions (DelSole and Tippett 2015) computed for that geographical box (Fig. 1). A multimodel ensemble (MME) is formed by averaging the individual model probabilities (Fig. 2) and L-ELR1–3 forecasts obtained by using

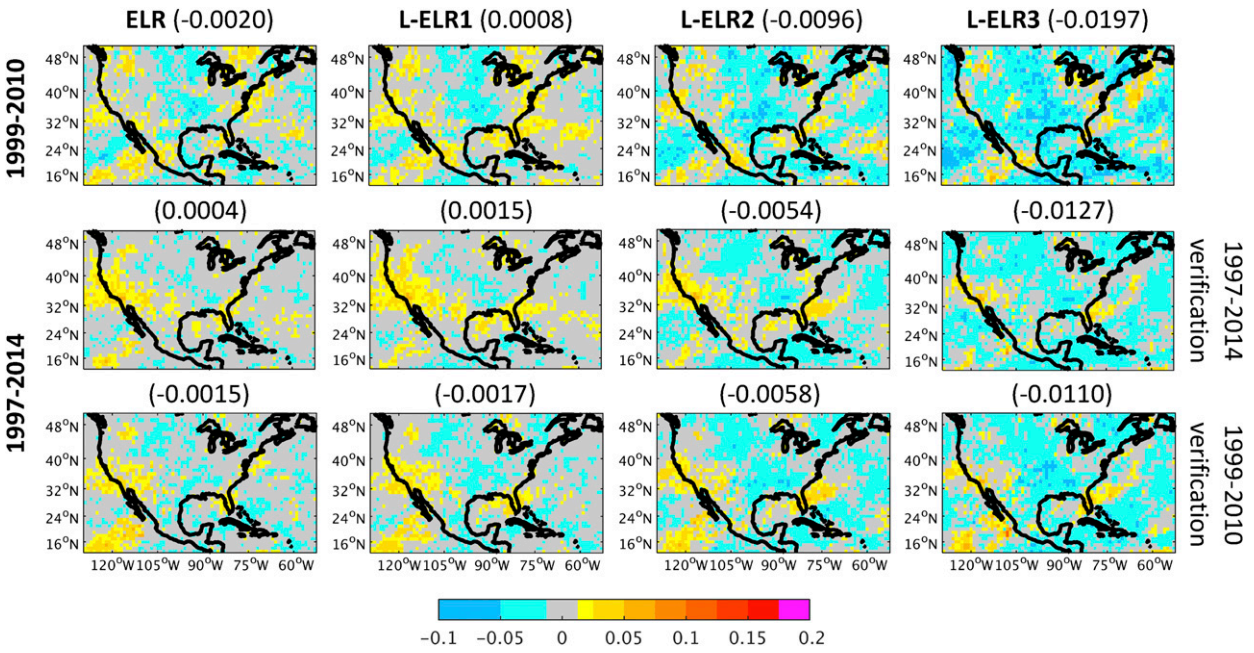

FIG. 13. RPSS for week-3–4 outlooks from ECMWF tercile precipitation ELR and L-ELR1–3 forecasts, for all starts during the JFM season when using reforecasts from the (top) 1999–2010 and 1997–2014 periods, when the latter are verified over (middle) 1997–2014 or (bottom) 1999–2010. Mean RPSS is indicated in parentheses for each forecast.

1 to 3 Laplacian eigenfunctions as predictors are compared to ELR forecasts based on the gridpoint mean (Vigaud et al. 2017a).

L-ELR1 weekly precipitation tercile forecasts exhibit low sharpness (Figs. 4 and 5) and skill decreasing with leads and from winter to summer (Figs. 6–8). However, skill is increased compared to ELR (Fig. 9) indicating more skillful predictions when using spatially averaged precipitation instead of the gridpoint ensemble mean as predictor. The size of this 15 × 15 gridpoint box used to compute Laplacian eigenfunctions is consistent with meteorological synoptic scales such as those of midlatitude depressions for instance (thousands of kilometers), which could explain the increase in winter skill when including spatial information. After week 2, reliability and resolution drop over North America, where skill remains low for all forecasts. Skill also decreases when including more Laplacians as additional explanatory variables. This can be explained by the sensitivity of regressions to sample variability, which increases with the number of predictors, leading to overconfident probability forecasts as reflected by high sharpness as lead increases for L-ELR2–3, and suggesting that improvements are limited by the small size of reforecasts used to train the ELR model.

Skill is enhanced from week-3 and week-4 forecasts when combining both leads to form week-3–4 tercile probabilities (Figs. 10 and 11). The 2-week targets are in line with the concept of seamless predictions (Zhu et al. 2014) and might have the advantage to capture better the time-scales of rainfall relationships to shifts in the jet stream and storm tracks, including those induced by ENSO and the MJO. Skill is maximized for L-ELR1 with almost 2 times the skill of week-3 leads (Fig. 12) and highest skill over the Northeast, West, and Southwest in JFM, and the western United States in JAS. These patterns of maximum skill are consistent with the maximum influence of jet-stream and storm-track modulations on North American rainfall in winter and the highest predictability of tropical rainfall in eastern Pacific regions related to ENSO in summer, respectively. L-ELR2–3 exhibit lower skill because of the small sample size. However, increased RPSS (Fig. 13) and reduced overconfidence (not shown) for ECMWF week-3–4 outlooks when verifying and extending the pool of reforecasts to 1997–2014 indicate potential for further skill improvements by increasing reforecasts length.

Skill relationships to large-scale tropical forcings such as ENSO or the MJO are maximized in L-ELR1 forecasts, with greater RPSS relative to ELR (Fig. 14). Highest skill in all forecasts for winter starts during El Niño and MJO phases 2–3 and 6–7 are consistent with El Niño and MJO modulations of the jet stream and U.S. precipitation. Even if skill

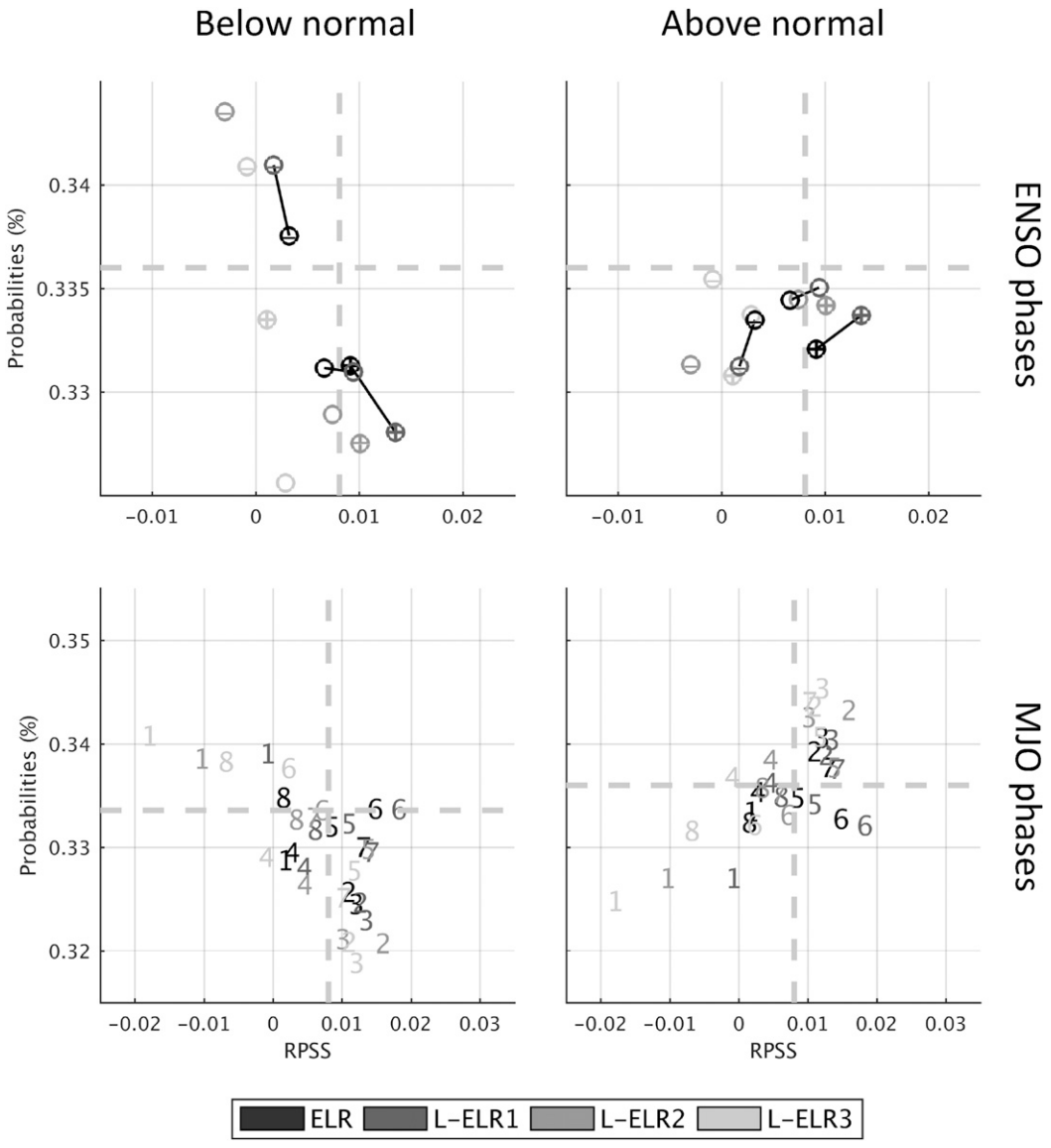

FIG. 14. Mean weekly RPSS vs (left) below- and (right) above-normal probabilities averaged over continental North America between 20° and 50°N latitudes, for ELR and L-ELR1–3 precipitation tercile MME forecasts with starts in JFM during observed (top) ENSO and (bottom) MJO phases. El Niño, neutral ENSO, and La Niña phases are indicated by ⊕, ○, and ⊖ symbols, respectively, and those of the MJO are indicated by their respective number. Dashed lines correspond to the 0.1 level of significance using a permutation test, and black lines in the top panels indicate the skill gain from ELR to L-ELR1.

remains low, these results suggest increased opportunities for skillful predictions through spatial pattern correction and increasing length of reforecast archives.

Acknowledgments. The authors thank the two anonymous reviewers for their insightful comments, which helped improving the manuscript substantially. The authors are grateful for 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 with IRI resources and this work is based on GPCP 1DD 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 R. Livezey, 1987: A high resolution rotated EOF analysis of monthly and seasonally averaged 700 mb heights. *Mon. Wea. Rev.*, **115**, 1083–1126, https://doi.org/10.1175/1520-0493(1987)115<1083:CSAPOL>2.0.CO;2.

——, 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.

Brier, G., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

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.

Dettinger, M., 2011: Climate change, atmospheric rivers and floods in California—A multimodel analysis of storm frequency and magnitude changes. *J. Amer. Water Resour. Assoc.*, **47**, 514–523, https://doi.org/10.1111/j.1752-1688.2011.00546.x.

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.

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.

Gneiting, T., and A. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.*, **102**, 359–378, https://doi.org/10.1198/016214506000001437.

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.

——, 2012: Verification of TIGGE multimodel and ECMWF reforecast-calibrated probabilistic precipitation forecasts over the contiguous United States. *Mon. Wea. Rev.*, **140**, 2232–2252, https://doi.org/10.1175/MWR-D-11-00220.1.

Huffman, G., and D. Bolvin, 2012: Version 1.2 GPCP One-Degree Daily (1DD) precipitation dataset documentation. NASA Goddard Space Flight Center Doc., 27 pp., http://apdrc.soest.hawaii.edu/doc/gpcp_daily.pdf.

——, R. Adler, M. Morrissey, D. Bolvin, S. Curtis, R. Joyce, B. McGavock, and J. Susskind, 2001: Global precipitation at one-degree daily resolution from multisatellite observations. *J. Hydrometeor.*, **2**, 36–50, https://doi.org/10.1175/1525-7541(2001)002<0036:GPAODD>2.0.CO;2.

Kharin, V., and F. Zwiers, 2003: Improved seasonal probability forecasts. *J. Climate*, **16**, 1684–1701, https://doi.org/10.1175/1520-0442(2003)016<1684:ISPF>2.0.CO;2.

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.

Monteverdi, J., and J. Null, 1998: A balanced view of the impact of the 1997/98 El Niño on Californian precipitation. *Weather*, **53**, 310–313, https://doi.org/10.1002/j.1477-8696.1998.tb06406.x.

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.

Ralph, F. M., P. J. Neiman, G. N. Kiladis, K. Weickmann, and D. W. Reynolds, 2011: A multiscale observational case study of a Pacific Atmospheric river exhibiting tropical-extratropical connections and a mesoscale frontal wave. *Mon. Wea. Rev.*, **139**, 1169–1189, https://doi.org/10.1175/2010MWR3596.1.

Robertson, A., U. Lall, S. E. Zebiak, and L. Goddard, 2004: Improved combination of multiple atmospheric GCM ensembles for seasonal prediction. *Mon. Wea. Rev.*, **132**, 2732–2744, https://doi.org/10.1175/MWR2818.1.

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: Data analysis and representation on a general domain using eigenfunctions of Laplacian. *J. Climate*, **28**, 7420–7436, https://doi.org/10.1016/j.acha.2007.09.005.

Smith, T., and R. Livezey, 1999: GCM 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.

van den Dool, H., 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.

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**, 67, 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. *Wea. Forecasting*, **33**, 1513–1532, https://doi.org/10.1175/WAF-D-18-0074.1.

Vitart, F., 2014: Evolution of ECMWF sub-seasonal forecast skill scores. *Quart. J. Roy. Meteor. Soc.*, **140**, 1889–1899, https://doi.org/10.1002/qj.2256.

——, 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.

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.

Zhang, C., 2013: Madden–Julian Oscillation: Bridging weather and climate. *Bull. Amer. Meteor. Soc.*, **94**, 1849–1870, https://doi.org/10.1175/BAMS-D-12-00026.1.

Zhu, H., M. Wheeler, A. Sobel, and D. Hudson, 2014: Seamless precipitation prediction skill in the tropics and extra tropics from a global model. *Mon. Wea. Rev.*, **142**, 1556–1569, https://doi.org/10.1175/MWR-D-13-00222.1.