Do Statistical Pattern Corrections Improve Seasonal Climate Predictions in the North American Multimodel Ensemble Models?

Anthony G. Barnston

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

Michael K. Tippett

Department of Applied Physics and Applied Mathematics, Columbia University, New York, New York, and Center of Excellence for Climate Change Research, Department of Meteorology, King Abdulaziz University, Jeddah, Saudi Arabia

(Manuscript received 31 January 2017, in final form 7 July 2017)

ABSTRACT

Canonical correlation analysis (CCA)-based statistical corrections are applied to seasonal mean precipitation and temperature hindcasts of the individual models from the North American Multimodel Ensemble project to correct biases in the positions and amplitudes of the predicted large-scale anomaly patterns. Corrections are applied in 15 individual regions and then merged into globally corrected forecasts. The CCA correction dramatically improves the RMS error skill score, demonstrating that model predictions contain correctable systematic biases in mean and amplitude. However, the corrections do not materially improve the anomaly correlation skills of the individual models for most regions, seasons, and lead times, with the exception of October–December precipitation in Indonesia and eastern Africa. Models with lower uncorrected correlation skill tend to benefit more from the correction, suggesting that their lower skills may be due to correctable systematic errors. Unexpectedly, corrections for the globe as a single region tend to improve the anomaly correlation at least as much as the merged corrections to the individual regions for temperature, and more so for precipitation, perhaps due to better noise filtering. The lack of overall improvement in correlation may imply relatively mild errors in large-scale anomaly patterns. Alternatively, there may be such errors, but the period of record is too short to identify them effectively but long enough to find local biases in mean and amplitude. Therefore, statistical correction methods treating individual locations (e.g., multiple regression or principal component regression) may be recommended for today's coupled climate model forecasts. The findings highlight that the performance of statistical postprocessing can be grossly overestimated without thorough cross validation or evaluation on independent data.

1. Introduction

In principle, dynamical climate prediction models are expected to produce more accurate climate predictions than statistical models on seasonal to interannual time scales. This expectation is based on the fact that dynamical models make use of the often complex and nonlinear physical laws governing oceanic and atmospheric behavior, while statistical models use only relationships (often linear) gleaned from finite records of observational data. Operationally, however, dynamical models did not show clear superiority over statistical models in predicting monthly or seasonally averaged climate until near the turn of the twenty-first century, as more advanced data assimilation methods and computer power finally enabled them to perform closer to their potential.

While comprehensive coupled ocean–atmosphere dynamical models are now heavily relied upon for seasonal climate predictions, they still have aspects in need of further improvement. Their systematic errors, or biases, vary by model, season, lead time, and location. The presence of biases creates an opportunity for statistical models to detect and correct them, resulting in improved final forecast quality. Such methods can be

Denotes content that is immediately available upon publication as open access.

Corresponding author: Anthony G. Barnston, tonyb@iri.columbia.edu

DOI: 10.1175/JCLI-D-17-0054.1

TABLE 1. Basic information for the eight models of the NMME used in the study. The acronym FLOR stands for Forecast-oriented Low Ocean Resolution.

Model	Expanded model name	No. ensemble members	Max lead (months)	Reference
1) CMC1-CanCM3	Canadian coupled model 1	10	12	Merryfield et al. (2013)
2) CMC2-CanCM4	Canadian coupled model 2	10	12	Merryfield et al. (2013)
3) COLA-RSMAS-CCSM4	COLA/University of Miami/NCAR coupled model	10	12	Gent et al. (2011) and Infanti and Kirtman (2016)
4) GFDL-CM2pl-aer04	Modified version of GFDL coupled model	10	12	Delworth et al. (2006) and Zhang et al. (2007)
5) GFDL-CM2p5-FLOR-A06	Expanded version of GFDL coupled model, FLOR-A06	12	12	Vecchi et al. (2014)
6) GFDL-CM2p5-FLOR-B01	Expanded version of GFDL coupled model, FLOR-B01	12	12	Vecchi et al. (2014)
7) NASA-GMAO-062012	Modified version of NASA coupled model	12	9	Vernieres et al. (2012)
8) NCEP-CFSv2	NOAA/NCEP coupled model	24	10	Saha et al. (2014)

used to modify the positions and/or amplitudes of large-scale patterns and also to refine the details of anomaly patterns for local downscaling. Here, we apply statistical corrections to the models in the North American Multimodel Ensemble (NMME) and focus on the correction of biases in the positions and amplitude of the models' predicted large-scale anomaly patterns.

The statistical treatment of systematic errors in the positions and amplitudes of patterns in dynamical model predictions is not new. Rukhovets et al. (1998) used singular value decomposition (SVD) to document the patterns of 500-hPa height fields in medium-range forecasts of a NOAA/National Centers for Environmental Prediction (NCEP) model and their correspondences to observed patterns that would enable a model pattern calibration. Ward and Navarra (1997) and Smith and Livezey (1999) used canonical correlation analysis (CCA) to relate seasonal climate predictions or simulations (with prescribed SST fields) of a general circulation model (GCM) to the corresponding observations in large regions of interest, with the potential to better calibrate the forecasts. Similarly, Feddersen et al. (1999) used SVD to postprocess precipitation predictions from a GCM, and Mo and Straus (2002) used principal component regression (PCR) with similar purpose for the climate in the Northern Hemisphere as well as smaller regions. Motivated by a severe drought in central southwest Asia during the extended 1998–2000 La Niña, Tippett et al. (2003) used CCA to correct GCM predictions of precipitation in that region. Corrections by CCA, SVD, or PCR differ primarily in the prefiltering of the datasets (Tippett et al. 2008). In all of the above studies, the GCM predictions were forced by either predicted or observed SST, in contrast to today's comprehensive dynamical model forecasts in which the SST and climate are predicted simultaneously in a single-tier coupled model integration.

In phase one of the NMME project, hindcasts were generated for global fields of SST, surface air temperature, precipitation, and other variables from eight or more state-of-the-art coupled GCMs (Kirtman et al. 2014). These hindcasts were for monthly average climate extending to up to 12 months into the future, spanning the 1982–2010 period. Real-time predictions from the same models began in 2011, adding to the data archive where the hindcasts ended.

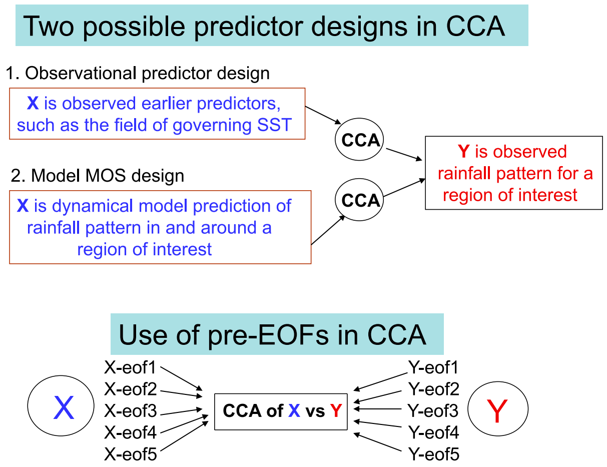
The purpose of this study is to determine whether the commonly used multivariate statistical method of CCA can improve the temporal anomaly correlation skill of the individual NMME models, with the goal of improving the predictions of the multimodel ensemble. The anomaly correlation is used as the primary metric because it measures the ability to reproduce the phasing of the interannual variability of the climate. A secondary verification measure, based on the mean squared error, is also examined to see if local systematic biases are also reduced by the CCA.

The data and the analysis methods used in the study are described in section 2, followed by results, first for forecasts of precipitation and then temperature, in section 3. Concluding remarks and discussion are found in section 4.

2. Data and methods

a. Data

The model data used here are hindcasts from eight models of the NMME, spanning 1982–2010. The models are listed in Table 1, along with some of their basic characteristics and references. (The monthly hindcast data for these models are available at http://iridl.ldeo.columbia.edu/SOURCES/.Models/.NMME.) The eight models include one from the Center for Ocean–Land–Atmosphere Studies (COLA) and the University of



Two possible predictor designs in CCA

1. Observational predictor design

X is observed earlier predictors, such as the field of governing SST

2. Model MOS design

X is dynamical model prediction of rainfall pattern in and around a region of interest

Y is observed rainfall pattern for a region of interest

Use of pre-EOFs in CCA

Fig. 1. Schematics showing (top) two common uses of CCA in atmospheric sciences, where the second one is used in this study and (bottom) reduction of the number of original variables (spanning up to thousands of grid points) for the predictors X and predictands Y to just their several leading independent principal components, using EOF analysis, prior to applying CCA.

Miami, one from the National Aeronautics and Space Administration (NASA), three from the Geophysical Fluid Dynamics Laboratory (GFDL), two from the Canadian Meteorological Centre (CMC), and one from NOAA's NCEP. The global hindcast data are on a 1-degree grid. The eight models used provide varying numbers of ensemble members, ranging from 10 to 24. Here, the ensemble mean is used to represent the forecast signal, while the ensemble member spread, representing the forecast uncertainty and making possible probability forecasts, is not considered.

The verifying observations, also in a 1-degree grid and available on the above-cited web page, are CMAP-Unified Raingauge Dataset (URD) for precipitation (Xie and Arkin 1997) and GHCN Climate Anomaly Monitoring System (GHCN-CAMS) (Fan and Van den Dool 2008) for temperature, respectively, both created at the NOAA Climate Prediction Center. We also test an alternative observed dataset for temperature, the CAMS (Ropelewski et al. 1984; available at http://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NCEP/.CPC/.CAMS/.anomaly/.temp/). Most of the attention in this study is devoted to precipitation prediction.

b. CCA

CCA has been used in two general ways in climate prediction (Fig. 1, top). In the first version, CCA is a purely statistical forecast model in itself, relating anomaly patterns in recent observations (e.g., sea surface temperature anomalies) to climate anomaly patterns in a subsequent season (e.g., precipitation anomalies) based on an extended hindcast period. The relationships are then used to make real-time forecasts. In this case no dynamical model is involved. Examples are seen in Barnett and Preisendorfer (1987), Barnston (1994), and Johansson et al. (1998), among others. The second version, used in this study, relates the raw outputs of dynamical model predictions to their corresponding observations for the targeted forecast time based on a hindcast period. Often the model output is for the same time and same field as the observations (e.g., both for precipitation), unless a different model predictor field is found to perform better (e.g., predicted geopotential height against observed precipitation). Also, the predictor domain is usually designed to be larger than the targeted domain, so that relevant features outside of the targeted domain can be used for better model error correction. In this second CCA usage, the CCA serves as a dynamical model postprocessor, correcting the model's systematic errors—a technique sometimes called model output statistics (MOS).

In the CCA method used here, preorthogonalization, using empirical orthogonal function (EOF) analysis, is done separately on the model hindcasts (the X variable, or predictor) and on the corresponding observations (the Y variable, or predictand), and a truncated set of the principal component time series from these EOFs is used as input to the CCA (Fig. 1, bottom). Preorthogonalization reduces the number of variables used by the CCA, reducing the potential for overfitting (Barnett and Preisendorfer 1987) while preserving the most coherent patterns of variability. In terms of cross-validated skill, using CCA with preorthogonalization has been found to be competitive with other available methods that identify coupled patterns (Bretherton et al. 1992). The EOF analysis can be done using correlations or covariances among the gridded values, and experiments are conducted to determine which choice leads to more effective model corrections. On average, covariance results in better CCA corrections for precipitation forecasts in our study, while correlation tends to be preferred for temperature forecasts. A likely reason for this outcome is discussed in section 4.

Within the CCA itself, a correlation matrix is first computed between the principal component time series of each predictor (predictand) and those of all predictands (predictors). Note that only cross-dataset pairings are used; no predictor–predictor or predictand–predictand correlation coefficients are used. This is generally not a square matrix, but becomes square and symmetric when post- or premultiplied by its transpose, and each operation is carried out for the predictor and predictand CCA solutions, respectively. The predictor

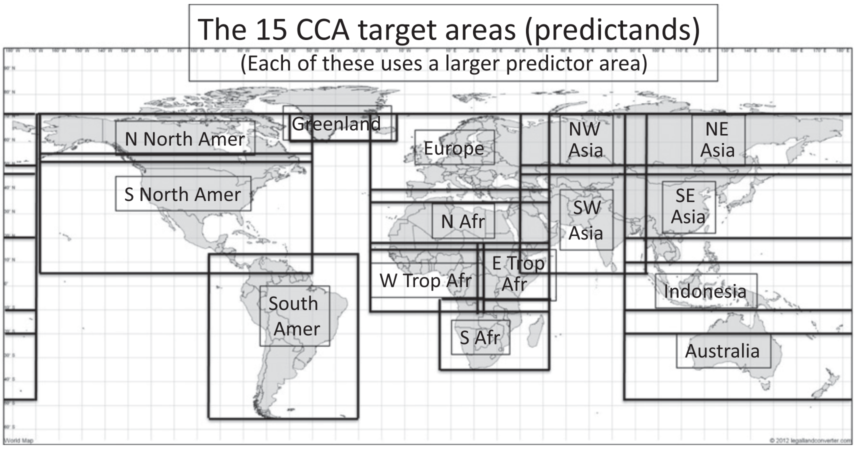


FIG. 2. The 15 slightly overlapping CCA target areas, each of which uses a larger predictor area.

and predictand product matrices are then used to obtain CCA eigenvalues and eigenvectors for predictors and predictands. The CCA modes consist of linear combinations of the predictor and predictand input principal components, which can be expanded back to geographical space through the ‘‘outer’’ EOFs used to preorthogonalize the initially larger datasets. Two CCA time series are also produced for predictor and predictand for each mode, and their correlation is called the canonical correlation coefficient. Hence, each mode shows a coupling of a predictor and predictand pattern, and the time series shows the strength and polarity with which each year showed the given pattern in the forecasts and observations. More detailed and complete mathematical descriptions of CCA are available in the appendix of Barnett and Preisendorfer (1987) and in Bretherton et al. (1992) and Tippett et al. (2008).

The CCA is applied to 15 different regions of the globe (Fig. 2), with the idea that each region is better treated with individual attention regarding the large-scale climate patterns pertinent to it but not necessarily to other regions of the globe. An attempt is made to define regions that capture coherent responses to known leading modes of variability, such as ENSO. Hence, eastern tropical Africa (E Trop Afr), southern Africa (S Afr), and southern North America (S North Amer) are used (Fig. 2). The regions overlap somewhat so that discontinuities in the forecasts near the boundaries may be smoothed using weighted averaging, where the weights are inversely related to the distance to the regions’ nearest borders. The corrected forecasts of each region are then merged to form a global forecast. The globe as a single region is also used as a 16th ‘‘region,’’ allowing for a skill comparison between the merged regional forecasts and the single globe forecast.

c. Cross validation

A cross-validation scheme is used in which three consecutive years are withheld from both the pre-EOF and the CCA training sample, and the middle year of the three is predicted. The years withheld progress from the earliest three to the latest three. The first and last years are also predicted so that each year has a cross-validated forecast. Three years, rather than just one, are withheld to minimize a negative skill bias that appears when there is a low correlation between forecasts and observations (Barnston and Van den Dool 1993). Also, withholding more than one year reduces a positive skill bias when there is substantial year-to-year autocorrelation of anomalies, as there are two available adjacent years in the hindcast tests when only one year is withheld but only one adjacent year (the previous year) in real-time forecasting. In many seasonal climate forecast settings, the negative bias has been shown to outweigh the positive one when the underlying skill level is modest.

The number of EOF modes used varies by region to approximately maximize skill, determined by cross-validated skill sensitivity tests that vary the numbers of modes for X, for Y, and for the CCA. Testing shows that for forecasts for a given region, season, and lead time, making the truncations model-specific adds little to resulting skill (less than 0.01 in anomaly correlation), and that within a range of truncation settings, corrected forecast skills for a given model are fairly insensitive to the setting. For example, when using an approximately optimum five EOF modes for X, six EOF modes for Y, and five CCA modes for a given model, truncations ranging from one less of each to two more for each result in little change in cross-validated skill.

d. Verification measures

The temporal anomaly correlation is used as the primary verification measure, and the average performance over a region is computed using the area-weighted average of the Fisher-transformed correlations.[1] The root-mean-square error skill score (RMSESS) is also computed to detect the presence of all types of calibration errors collectively—both forecast biases of the mean and the amplitude at the local level, as well as biases in the placement and amplitude of large-scale patterns.[2] The RMSESS is defined as

$$\text{RMSESS} = 1 - \frac{\sqrt{\text{MSE}_{\text{fct}}}}{\sqrt{\text{MSE}_{\text{clim}}}}, \qquad (1)$$

where MSE_{fct} is the MSE of the model forecasts and MSE_{clim} is the MSE of climatology forecasts (always forecasting the climatological mean). The RMSESS is a variation of the MSESS (Murphy and Epstein 1989), which does not take the square roots of the MSE terms in Eq. (1). When MSE_{fct} is equal in size to MSE_{clim}, RMSESS (and MSESS) is zero, and when MSE_{fct} is zero (i.e., all perfect forecasts), RMSESS (and MSESS) is 1. Note that $\sqrt{\text{MSE}_{\text{clim}}}$ equals the interannual standard deviation of the observations for a given season and location. The MSE_{fct} used in the RMSESS is based on standardized anomalies with respect to the observed mean and standard deviation.

e. Statistical significance

The statistical significance of improvements to the anomaly correlation is assessed using the Fisher Z transform (Hays 1973). For regional average correlations, an estimate of the number of spatial degrees of freedom in the region is made using the methods described in Van den Dool and Chervin (1986) and Moron et al. (2007). The degrees of freedom used in the Fisher Z test are then the product of the temporal degrees of freedom (based on 29 years of data) and the estimated spatial degrees of freedom. The Fisher test assumes independent samples from which the difference between the two correlations is compared; however, in our case the two samples are not independent because both share the same observation dataset. To account for this dependency, the method described in DelSole and Tippett (2014) is used [see their Eqs. (11) and (12)], which removes the effect of the sample dependency by adjusting (usually decreasing) the standard error of the difference between the two transformed correlations based on the two correlation skills in conjunction with the correlation between the uncorrected and the corrected forecasts. Note that the approach to assessing the statistical significance for an entire region just described is an analytic alternative to the Monte Carlo approach to field significance put forth in Livezey and Chen (1983).

Another aspect of statistical significance applies to maps showing the spatial distribution of changes in the correlation skill due to the CCA correction. We want to know at which locations positive changes are statistically significant. While significance of improvements at individual grid points is assessed using the Fisher Z test described above, we also need to account for the multiplicity of significance tests being conducted across the entire map—for example, we expect about 5% of the locations to attain significance at the 0.05 level due purely to chance. This multiplicity issue is addressed here using the approach of the "false discovery rate" (i.e., the rate of making a type I statistical error), documented in Benjamini and Hochberg (1995) and applied to the atmospheric sciences in Wilks (2006, 2016). In this approach, the significance p values for all of the grid points are rank ordered from smallest to largest. Then each respective p value is compared to the quantity (0.05) (rank/total) where the smallest p value has rank 1, next smallest rank 2, and so forth, and "total" is the number of locations in the entire domain being evaluated. The value 0.05 is the false discovery rate control, which can be set to other values if desired. Only locations whose p values are no larger than the largest one that is less than 0.05 (rank/total) are considered locally significant using this approach. The number of points passing the test is typically considerably smaller than the number originally attaining 0.05 significance.

Significance tests for differences in RMSESS use the F test for the ratio of two variances, those variances here being the MSE_{fct} [as used in Eq. (1)] of each of the two sets of forecasts. As in the case of tests for changes in the anomaly correlation, the temporal degrees of freedom is multiplied by the estimated spatial degrees of freedom, and the effect of nonindependent samples (since both share the same observations) is addressed in accordance with DelSole and Tippett (2014), where the required F value to attain significance is decreased as a function of the correlation between the errors of the uncorrected and the corrected forecasts [see their Eq. (10)].

[1] The Fisher Z equivalents to correlation coefficients (Hays 1973) can be linearly averaged, and the average then transformed back to a correlation, while the correlations themselves should not be linearly averaged.

[2] Although pattern biases often encompass local biases, some local biases may be specific to individual grid points or confined to small areas and unable to survive the truncated pre-orthogonalization and the CCA.

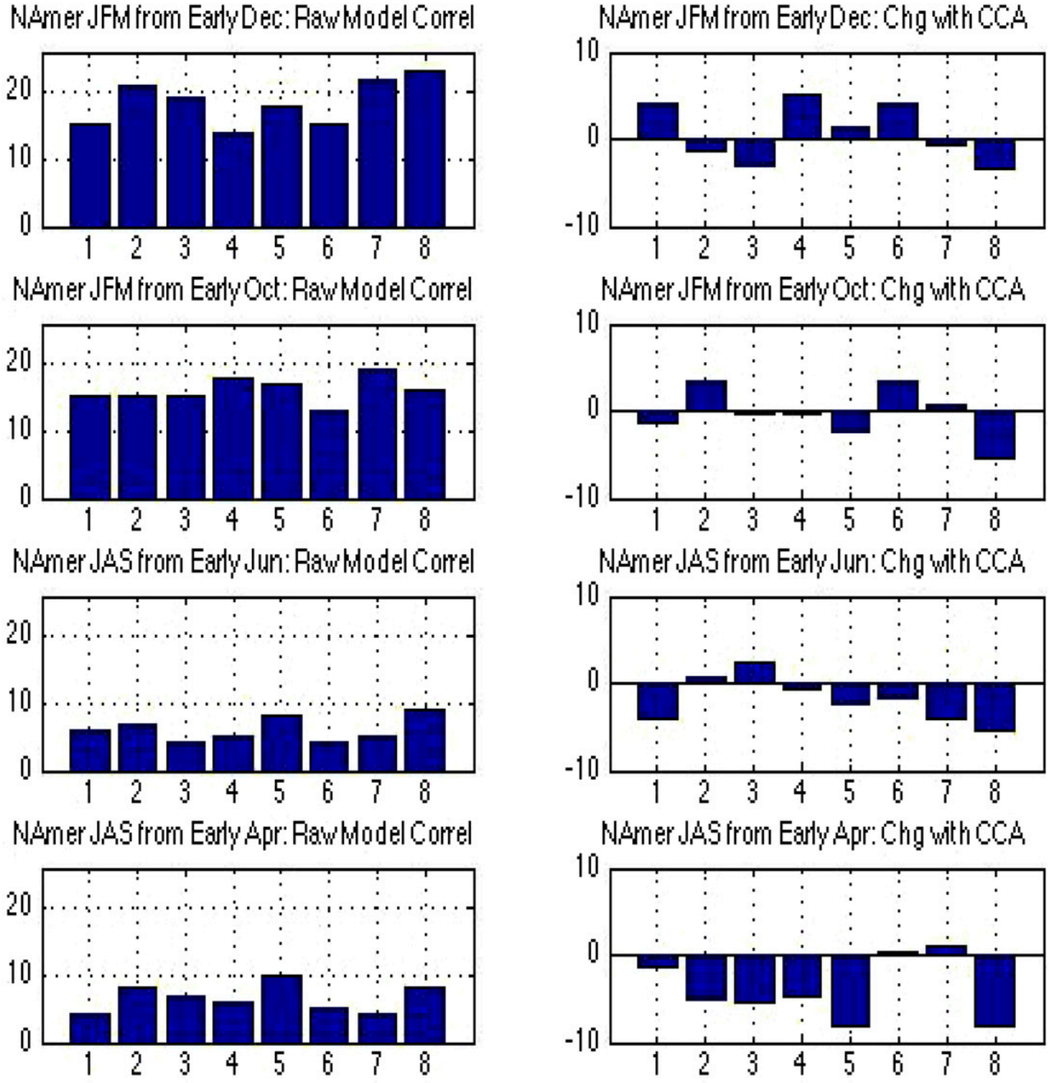


FIG. 3. (left) Original anomaly correlation skill (×100) and (right) the change in skill due to the CCA for the southern North America region for each of the eight NMME models for precipitation. The results (from top to bottom) are for (row 1) January–March precipitation forecasts from early December, (row 2) January–March forecasts from early October, (row 3) July–September forecasts from early June, and (row 4) July–September forecasts from early April. The order of the eight models (horizontal axis) is 1) CCSM4, 2) NASA, 3) GFDL, 4) GFDL-FLOR-A, 5) GFDL-FLOR-B, 6) CMC1, 7) CMC2, and 8) CFSv2.

All significance levels given here are for one-sided tests because of the a priori expectation that the CCA improves skill by accounting for systematic model errors.

3. Results

a. *Seasonal precipitation*

For CCA treatment of the precipitation forecasts, the covariance matrix is found to result in higher average skill improvement than the correlation matrix when used in the EOF analyses preceding the CCA. Therefore, it is used for all of the precipitation corrections.

Figure 3 shows, for the southern North America region (including the United States) for each of the eight models, the original area-averaged anomaly correlation skill and the change in skill due to the CCA for precipitation forecasts for the January–March and July–September target seasons. Skill for each of the target seasons is shown when predicted at lead times of 1.5 months (e.g., a January–March forecast made in early December) and 3.5 months (made in early October). Original model skill is approximately 0.15 to 0.20

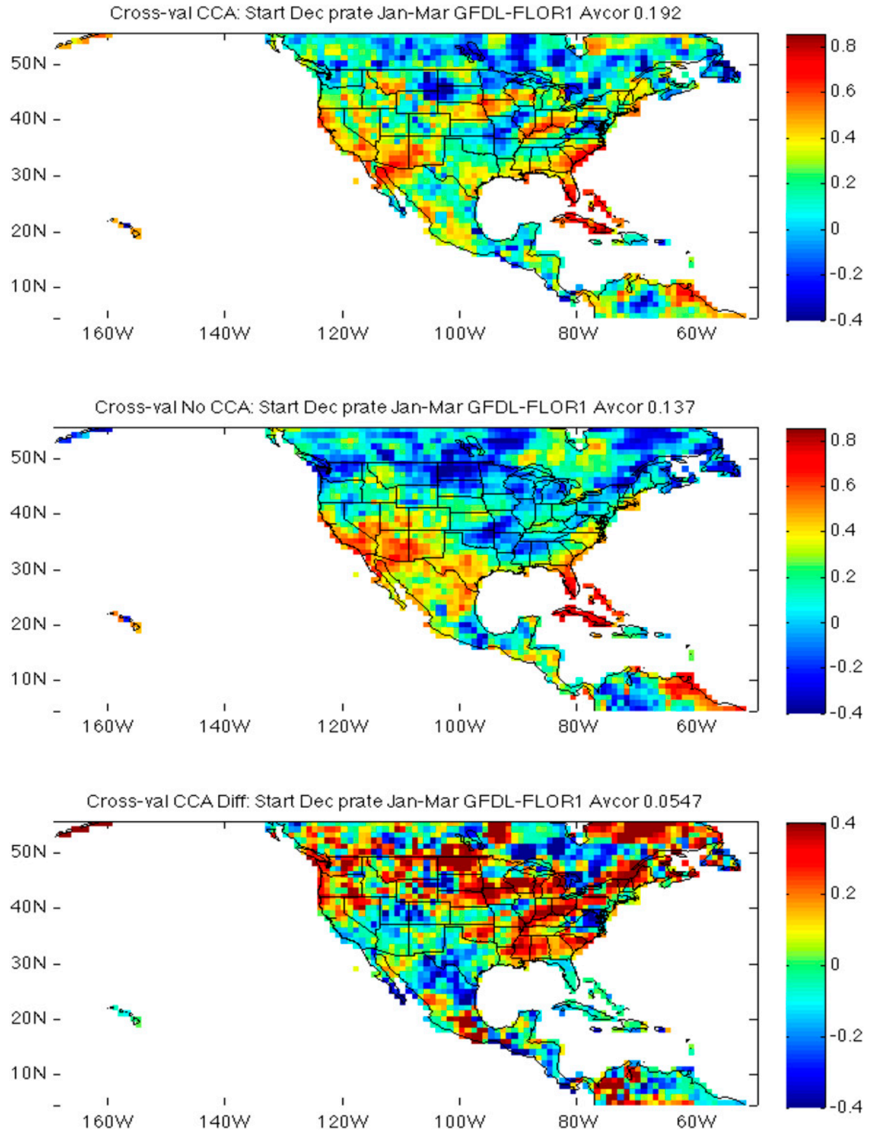

Cross-val CCA: Start Dec prate Jan-Mar GFDL-FLOR1 Avcor 0.192
Cross-val No CCA: Start Dec prate Jan-Mar GFDL-FLOR1 Avcor 0.137
Cross-val CCA Diff: Start Dec prate Jan-Mar GFDL-FLOR1 Avcor 0.0547
50N
40N
30N
20N
10N
160W
140W
120W
100W
80W
60W
0.8
0.6
0.4
0.2
0
-0.2
-0.4

FIG. 4. Geographic distribution of temporal anomaly correlation skill over the southern North America region for precipitation forecasts by the GFDL-FLOR-A model for January–March made in early December. (top) The skill after the CCA correction, (middle) the original skill, and (bottom) the skill improvement due to the CCA [note the different scale for (bottom)].

for January–March, with little drop in skill between early December and early October starts. Skill is less than 0.10 for most July–September forecasts. For the January–March forecasts from early December, the CCA corrections increase skill for the CCSM4, GFDL-FLOR-A, and CMC1 models, but result in little change or a decrease in skill for the other models. The corrections have similar effects on the longer lead forecasts for January–March, whereas for July–September forecasts the results of the corrections are mainly unfavorable. Overall, with the exception of a few cases, the CCA corrections do not result in substantial winter or summer skill improvements for precipitation forecasts in the southern North America region.

Although Fig. 3 does not show general improvement in skill, some of the models do show a skill increase for January–March forecasts made in early December (upper right panel). The skill of the GFDL-FLOR-A model (model 4) is improved by about 0.05. To further detail this result, Fig. 4 shows the spatial distribution of the anomaly correlation skill before and after the CCA correction and the skill change due to the CCA. Skill is improved in some

TABLE 2. Uncorrected anomaly correlation skill and the change in skill due to the CCA for precipitation forecasts for January–March made in early December, averaged over eight models, for each of 15 individual regions and for the globe as a single region. The first column shows the number of modes retained for the *X* EOFs, *Y* EOFs, and the CCA, followed by the model-average percentage of variance preserved after the EOFs of *X* and *Y*. The area-weighted average change in skill of the 15 individual regions is −0.02.

Region	No. modes *X*, *Y*, CCA; %Var *X*, *Y*	Initial skill	CCA: Skill change	Region	No. modes *X*, *Y*, CCA; %Var *X*, *Y*	Initial skill	CCA: Skill change
N North America	6, 7, 6 79, 53	0.05	0.03	South Africa	5, 5, 5 60, 53	0.10	0.04
S North America	6, 7, 6 84, 62	0.18	0.01	NW Asia	6, 7, 6 85, 74	0.10	0.06
South America	6, 7, 5 79, 50	0.15	−0.04	SW Asia	6, 7, 6 75, 63	0.13	−0.06
Greenland	5, 5, 4 69, 76	0.06	0.05	NE Asia	6, 7, 6 69, 58	0.09	−0.01
Europe	6, 7, 6 67, 60	0.07	−0.05	SE Asia	6, 7, 6 78, 62	0.12	−0.06
North Africa	5, 6, 5 61, 54	0.07	−0.04	Indonesia	5, 4, 3 76, 67	0.24	0.01
W Trop Africa	6, 7, 6 65, 63	0.01	−0.01	Australia	5, 6, 4 80, 67	0.24	−0.14
E Trop Africa	5, 5, 5 68, 65	0.05	−0.16	Single Globe	16, 18, 16 84, 80	0.114	0.000

portions of the United States (the Midwest, northern Plains, and Pacific Northwest) but degraded in other portions of the domain. Using the Fisher *Z* transform to assess the statistical significance of the positive skill changes at each grid point, 20% of the points have significant positive changes at the 0.05 level and 12% at the 0.01 level. However, only 9% are significant using the false discovery rate approach, using 0.05 as the false discovery rate control level. This would mean that only small portions of the darkest red locations on the bottom panel of Fig. 4 are locally significant.[3] To evaluate the 0.05 domain-average skill improvement (from 0.14 to 0.19), we estimate 5 spatial degrees of freedom for precipitation for this region during winter and use the dependent-sample version of the Fisher *Z* test, and find that the improvement in correlation is not statistically significant at the 0.10 level. This result implies that none of the domain-average improvements shown in the panels on the right side of Fig. 3 are statistically significant and that the variation of CCA-related changes among the models due to the CCA could be largely due to sampling variations rather than physically based differences in the models' responses to the linear pattern corrections. It should also be kept in mind, however, that with only 29 years of data, the power of statistical tests for the difference between two correlation skills is modest, and large differences are required to achieve statistical significance. In other words, part of the differences that fail the statistical significance test may still be real but are embedded in too much sampling variability for that variability to be considered sufficiently unlikely to have caused the result.

Despite this statistically negative result, a feature worth noting is that models having relatively higher uncorrected skill tend to be helped less by the CCA than models having lower starting skill. This might be the case if there is an upper limit of skill that is approximately the same from one model to another and if models farther from that level are more able to benefit from the CCA than those closer to it.

When averaged over the eight models, the change in skill from the CCA correction in southern North America for forecasts of January–March made in early December is 0.01. Table 2 shows the model-average skill changes for January–March forecasts for this short lead time for each of the 15 regions and indicates that southern North America is one of 6 regions out of 15 to have a positive net skill change, the highest of which occurred in northwestern Asia. However, none of the six regions' average skill improvements are statistically significant at the 0.10 level. Table 2 also shows the number of EOF modes retained in the preorthogonalization of *X* and *Y* and the percentages of original variance preserved in the process.

[3] Strictly speaking, the significant points would not be points simply exceeding a correlation improvement threshold because the original and corrected correlations themselves also matter (e.g., a change from 0.3 to 0.6 is more significant than a change from 0.0 to 0.3).

Table 3. Comparison of the effect on globally averaged anomaly correlation skill of the CCA when performed on individual regions and merged to a global precipitation forecast and when performed on the globe as a single region. Results are averaged over all eight models and are shown for forecasts for January–March made in early December and early October and forecasts for July–September made in early June and early April.

Precipitation start → target	Original model skill	Style	Change from CCA
Dec → JFM	0.114	Merge	−0.023
		Single globe	0.000
Oct → JFM	0.084	Merge	−0.008
		Single globe	0.017
Jun → JAS	0.086	Merge	−0.013
		Single globe	0.009
Apr → JAS	0.065	Merge	−0.007
		Single globe	0.017

The EOFs capture roughly two-thirds of the variance for most of the regions and generally retain more variance in the model predictions X than in the observations Y despite the larger areas covered in the former.

The CCA can also statistically calibrate the forecasts for the entire globe as a single region, rather than merging the corrected forecasts of the individual regions. One might expect the merged skill result to be better than the single globe result due to the individualized focus provided to each region when treated separately. However, this expectation is not confirmed. As shown in Table 2, the area-weighted average of the CCA-related skill change over the individual regions (−0.02) is slightly lower than the skill of the globe as a single region, which is near zero. The percentage of variance retained in the preorthogonalization EOFs for the single globe analysis is somewhat larger than that retained in most of the individual regions, even though cross-validated skill was optimized in each case. More will be said about this outcome in the final section.

Skill comparisons for other seasons and lead times generally give similar results to those for the short-lead precipitation forecasts for January–March, in that substantial skill improvements due to the CCA are in a minority, and the CCA for the single globe produces slightly better results than the individually tailored CCAs for each region and merged into a global forecast. A summary of these precipitation results is shown in Table 3 for the target seasons of January–March and July–September, each at 1.5- and 3.5-month lead times.

Exceptions to the unimpressive results shown above are found in a few specific regions and seasons. Predictions for the October–December season made in early September are more favorable in the case of Indonesia (Fig. 5) and to a lesser degree in eastern equatorial Africa (Fig. 6). In both of these regions—particularly Indonesia—the precipitation during October–December is significantly related to the ENSO state. Indonesia is in the western portion of the ENSO phenomenon itself (Walker and Bliss 1934; Bjerknes 1969), while eastern equatorial Africa has an ENSO teleconnection mediated by the SST anomaly in the western Indian Ocean (Goddard and Graham 1999). In these clear cases of inherent seasonal climate predictability, the historical rainfall observations add value to the models' already relatively skillful predictions. Statistical assessment shows that the skill change for the CFSv2 model in the Indonesia case and the CMC2 and NASA models in the east tropical Africa case are significant at the 0.10 level, while the other models fall short of this significance level.

Figure 7 shows the original individual model correlation skills and the change in skill due to the CCA for the globe as a single region, for forecasts of precipitation for January–March and for July–September, each made early in the preceding month as well as two months earlier. Forecasts for January–March from both December and October start times begin with a global average skill mostly near 0.1. After applying CCA, skills change slightly, becoming either lower or higher than their starting level. These skill changes are not statistically significant. Nonetheless, some features of the skills are noteworthy. Models showing a lower starting skill more frequently have a positive change from the CCA than those with a higher starting skill. The shorter lead forecasts for January–March show little average skill change with the CCA, while six out of eight models have a small positive skill change in the longer lead forecasts, with the CMC1-CanCM3 model skill being helped the most. Forecasts of July–September start with lower original skills than found for January–March, averaging just under 0.1 for early June starts and slightly more than 0.05 for early April starts. For early June starts there is little average skill change with the CCA, while for early April starts, seven out of eight models show slight improvements, with the CMC2-CanCM4 benefiting the most from the CCA. Estimating 23 spatial degrees of freedom for global precipitation in January–March, only the two cases of correlation skill improvement of 0.05 or greater in Fig. 7 are statistically significant at the 0.10 level; interestingly, both of these appear for the longer-lead forecasts.

The geographical distribution of correlation skill before and after the CCA, as well as the CCA-related skill change, for precipitation forecasts for January–March made in early December by the CMC1-CanCM3 model is shown in Fig. 8 for the globe treated as a single region. Global average skill is 0.09 and 0.12 before and after the

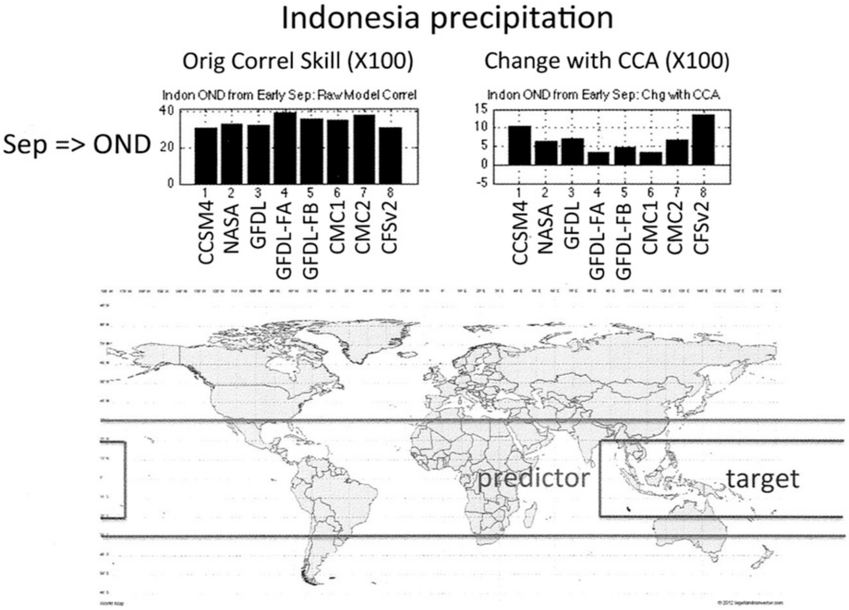


FIG. 5. (top),(left) Original anomaly correlation skill (×100) and (right) the change in skill due to the CCA for the Indonesia region for each of the eight NMME models. Results are for October–December precipitation forecasts from early September. (bottom) The location of the Indonesia forecast target region (land only) and its larger predictor region spanning the global tropics.

CCA, respectively. Improvements are seen in parts of tropical Africa, extratropical Asia, and North America. However, pockets of skill degradation appear in Australia, eastern equatorial Africa, and other smaller areas throughout the globe. These locations of skill change are likely partly real and partly due to sampling variability. The global average skill increase of 0.03 is not statistically significant.

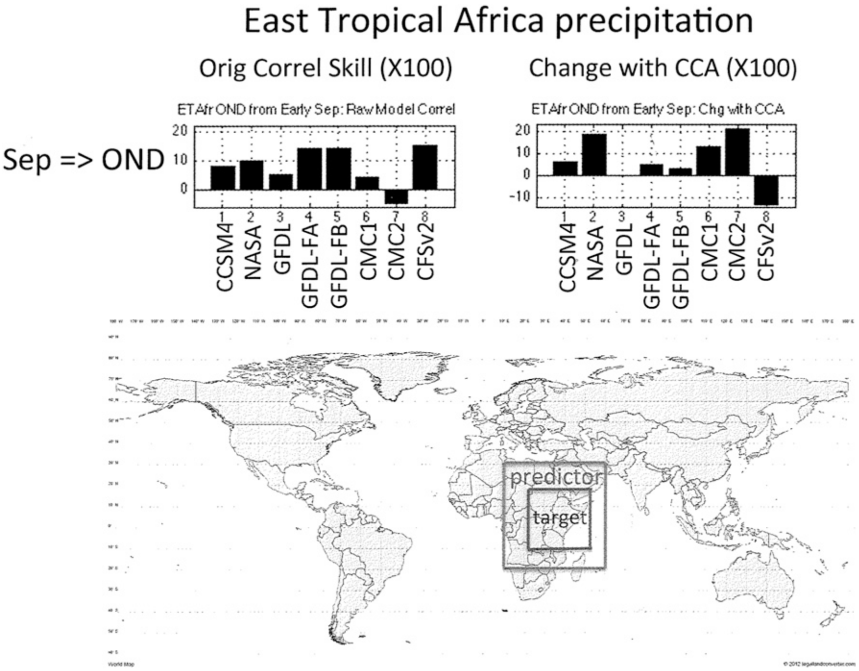


FIG. 6. (top)As in Fig. 5, but for the east tropical Africa region. (bottom) The location of the east tropical Africa forecast target region (land only) and its larger embedding predictor region.

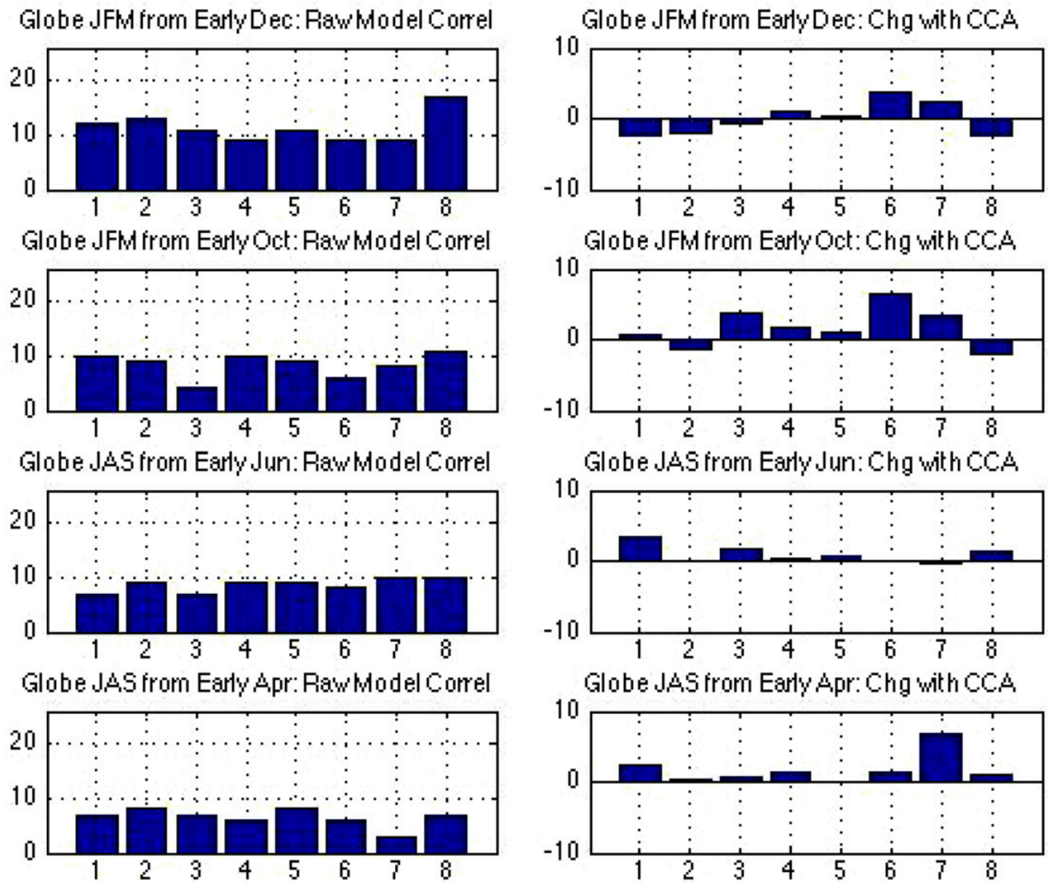
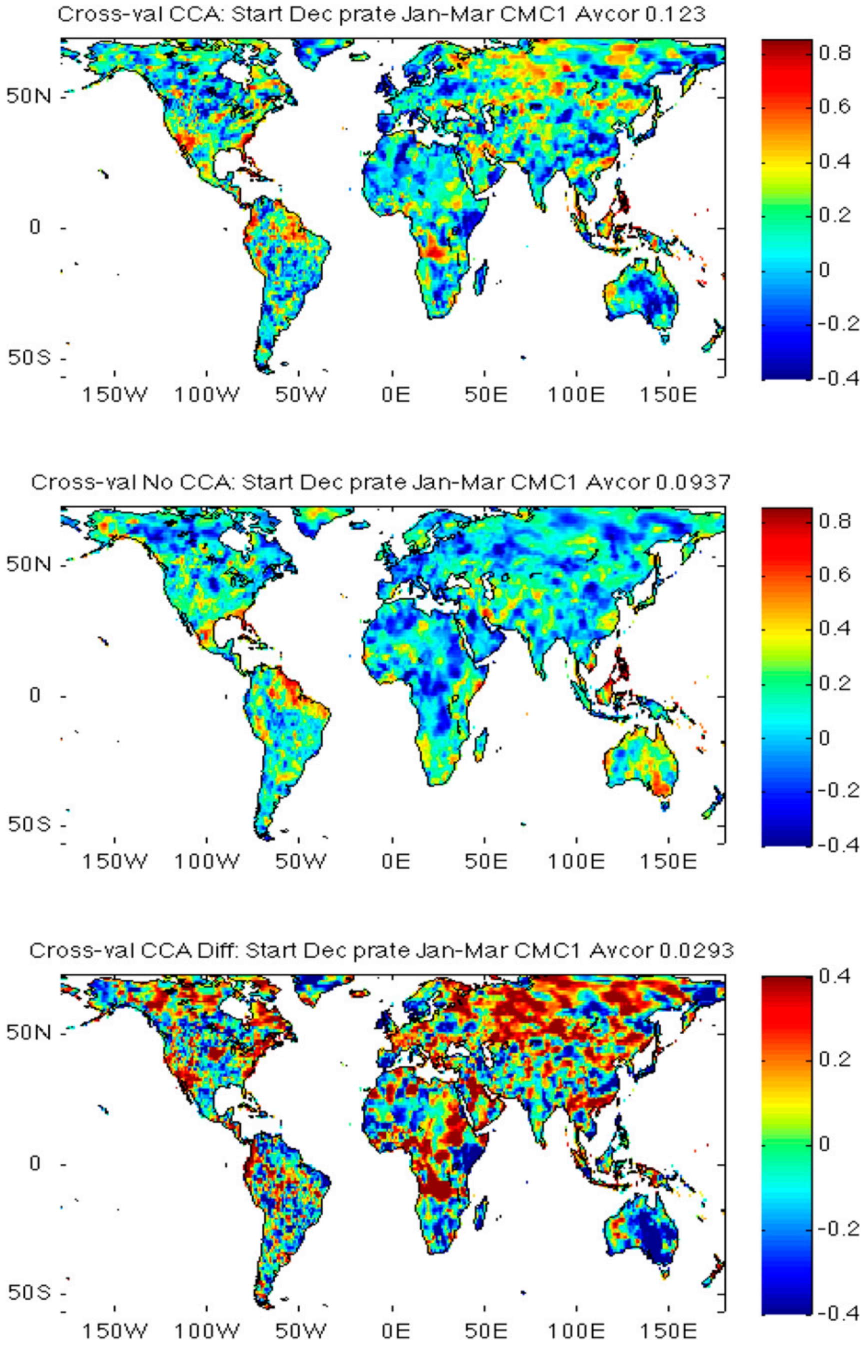


FIG. 7. (left) Original anomaly correlation skill (×100) and (right) the change in skill due to the CCA for the globe treated as a single region for each of the eight NMME models for precipitation. The results (from top to bottom) are for (row 1) January–March forecasts from early December, (row 2) January–March forecasts from early October, (row 3) July–September forecasts from early June, and (row 4) July–September forecasts from early April. The order of the eight models (horizontal axis) is 1) CCSM4, 2) NASA, 3) GFDL, 4) GFDL-FLOR-A, 5) GFDL-FLOR-B, 6) CMC1, 7) CMC2, and 8) CFSv2.

As an additional verification measure, the RMSESS between the precipitation forecasts and the corresponding observations is computed before and after the CCA correction. Figure 9 shows RMSESS of uncorrected and CCA-corrected forecasts of the globe treated as a single region, for the January–March season made from early October. In contrast to the anomaly correlation, the CCA substantially improves the RMSESS over most of the globe. Table 4 shows RMSESS for uncorrected and corrected forecasts for January–March and July–September, each at 1.5- and 3.5-month lead times. In all cases, considerable improvements in RMSESS are noted. The statistical significance of the difference between uncorrected and corrected RMSESS is tested using the F ratio of the uncorrected to the corrected $\mathrm{MSE}_{\mathrm{fct}}$ variance terms in Eq. (1). Most of the individual grid points are significant at $p < 0.01$ for the four season/lead time combinations shown in Table 4, and field significance for the entire map is still stronger owing to the additional spatial degrees of freedom. However, the average RMSESS of the corrected forecasts still falls just short of zero, meaning that it is still slightly larger than that of perpetual climatology forecasts, despite the positive global average correlation skills of the forecasts seen earlier in both the uncorrected and CCA-corrected forecasts.

The large improvement in RMSESS suggests the presence of systematic forecast errors in the uncorrected forecasts that do not necessarily involve spatial pattern placement. Such errors include mean biases and amplitude biases that are local, or of small spatial scale, and are generally not associated with the locational aspects of large-scale anomaly patterns. Correction of large-scale patterns would result in improved anomaly correlations as well as RMSESS. On the other hand, correction of purely local biases in forecast mean or



Cross-val CCA: Start Dec prate Jan-Mar CMC1 Avcor 0.123

Cross-val No CCA: Start Dec prate Jan-Mar CMC1 Avcor 0.0937

Cross-val CCA Diff: Start Dec prate Jan-Mar CMC1 Avcor 0.0293

FIG. 8. As in Fig. 4, but for the globe as a single region, for precipitation forecasts by the CMC1-CanCM3 model.

amplitude would improve RMSESS, as the corrected forecasts would have smaller errors with respect to the observations, but the anomaly correlation would not be changed because the temporal phasing of the forecasts and observations would be preserved. As the CCA is capable of improving both types of calibration problems with the appropriate truncations of the EOF and CCA modes, it appears that spatial pattern placement errors are only prominent in the case of some of the models for some seasons and regions, as shown by those instances in which CCA materially improves the correlation as well as RMSESS. Inspection of the NMME forecast data reveals numerous examples of substantial differences between forecast and observed means as well as

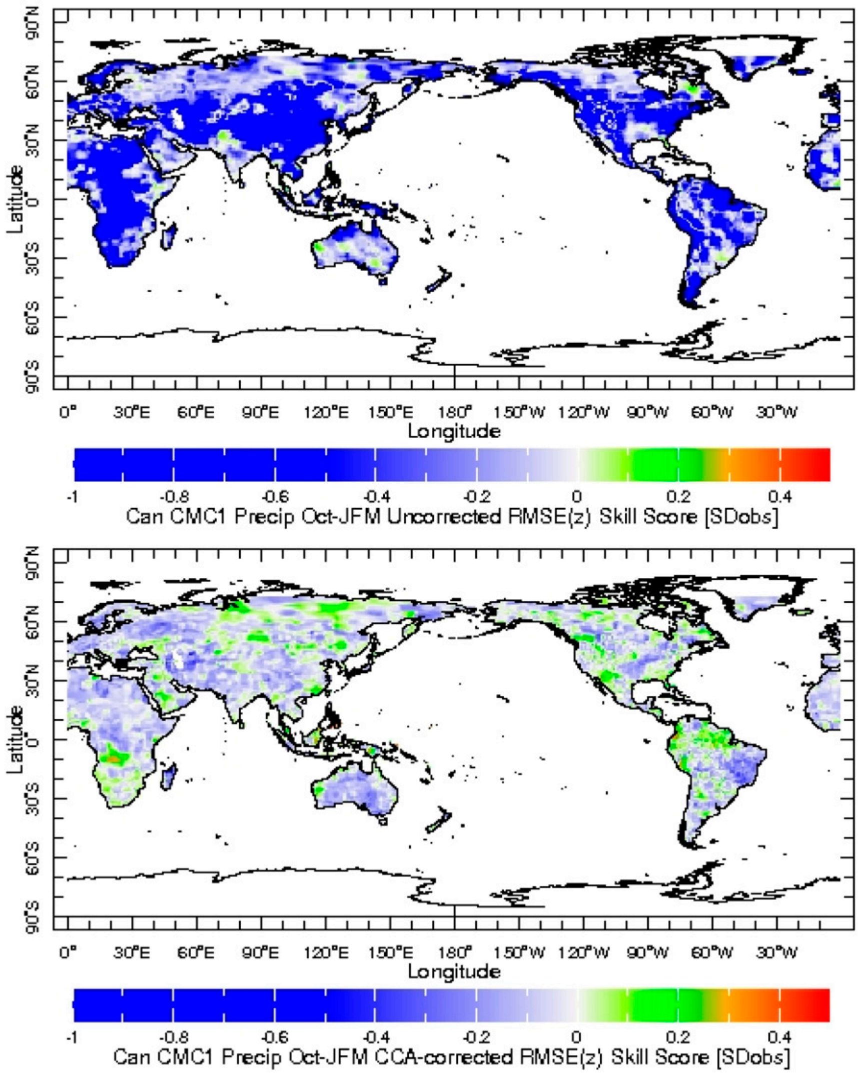

90°N
60°N
30°N
0°
30°S
60°S
90°S
Latitude
0°
30°E
60°E
90°E
120°E
150°E
180°
150°W
120°W
90°W
60°W
30°W
Longitude
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
Can CMC1 Precip Oct-JFM Uncorrected RMSE(z) Skill Score [SDobs]
Can CMC1 Precip Oct-JFM CCA-corrected RMSE(z) Skill Score [SDobs]

FIG. 9. Geographic distribution of root mean squared error skill score (RMSESS) over the globe as a single region, for precipitation forecasts by the CMC1-CanCM3 model for January–March made in early October. The top panel shows the original skill, and the bottom panel the skill following the CCA correction. The RMSESS is in terms of standardized anomalies with respect to the observed mean and standard deviation.

standard deviations (not shown). These appear coherent within small distances but otherwise vary in direction and size and are sometimes related to land–ocean interface or major terrain features.

If systematic errors in the positions of the large-scale anomaly patterns do not play a major role in the systematic errors of the NMME models, methods that focus on individual locations one at a time, such as multiple regression or principal component regression, may be adequate and simpler approaches, though requiring probabilistic reliability introduces challenges (Tippett et al. 2014). Here, the CCA has functioned to reduce systematic errors both locally and at the pattern level in the precipitation forecasts, but it might be compromised in handling the local biases compared with regression methods that do not filter the predictand data using the truncated sets of EOF and CCA modes.

b. *Seasonal temperature*

In contrast to precipitation, for temperature the correlation matrix is found to result in higher average skill improvement than the covariance matrix when used in the EOF prefiltering preceding the CCA. Therefore, the correlation matrix is used for all of the temperature corrections. When the CCA corrections are applied to seasonal temperature forecasts, a result approximately

TABLE 4. Global average RMSESS for precipitation before and after the CCA correction. Results are shown for forecasts for January–March made in early December and early October, and forecasts for July–September made in early June and early April.

Precipitation Start → Target	Global Avg RMSESS Before CCA	Global Avg RMSESS After CCA
Dec → JFM	−1.31	−0.04
Oct → JFM	−1.32	−0.04
Jun → JAS	−1.17	−0.05
Apr → JAS	−1.15	−0.05

similar to that of precipitation is found, but slightly less favorable. Prior to the CCA correction, forecasts for temperature usually have higher correlation skills than precipitation. However, as found for precipitation forecasts, the CCA only improves upon those skills in specific regions and seasons but in fewer cases than found for precipitation.

Table 5 summarizes the initial skill and the CCA-related skill change for the temperature forecasts of the 15 regions and of the globe as a single region, for forecasts of January–March made in early December. Most regions' skills are not improved by the CCA, and the average over all regions is negative. Also, in contrast with the result for precipitation, the CCA for the globe as a single region is not more favorable than the result for the 15 regions merged to form global forecasts. Table 5 shows the number of EOF modes retained in the pre-orthogonalization of X and Y and the percentages of original variance preserved in the process. The EOFs capture more than 80% of the variance for most of the regions, which exceeds that for precipitation (Table 2) even with fewer modes retained than for precipitation, as temperature has more spatial coherence (less noise) in its variability than precipitation.

While results averaged over all models are not favorable for most regions or for the globe as a single region, some models' temperature forecasts benefit from the CCA correction. For example, Fig. 10 shows the spatial distribution over the globe of correlation skill before and after the correction, along with the change due to the CCA, for forecasts of January–March made in early December by the GFDL-CM2p1 model. The initial model skill is good over parts of most continents and is improved most notably over eastern Europe/western Asia and in other regions. Skill is degraded by the CCA over Australia, northern Mexico, and some other regions. Global average correlation skill is increased from 0.28 to 0.32. However, estimating 15 spatial degrees of freedom for January–March temperature, the average correlation boost of 0.04 falls slightly short of statistically significance at the 0.10 level, even though 35% of the grid points are significant at the 0.05 level, 25% at the 0.01 level, and 27% are significant using the false discovery rate approach, using 0.05 as the false discovery rate control level. The significance of the average improvement is clearly hampered by regions having very strong degradations (e.g., Australia) even though only 40% of grid points had skill decreases.

A summary of results is shown in Table 6 for the target seasons of January–March and July–September, each at 1.5- and 3.5-month lead times. In forecasts for

TABLE 5. Uncorrected anomaly correlation skill, and the change in skill due to the CCA, for temperature forecasts for January–March made in early December, averaged over 8 models, for each of 15 individual regions and for the globe as a single region. The first column shows the number of modes retained for the X EOFs, Y EOFs, and the CCA, followed by the model-average percentage of variance preserved after the EOFs of X and Y. The area-weighted average change in skill of the 15 individual regions is −0.07.

Region	No. modes: X, Y, CCA; %Var X, Y	Initial skill	CCA: Skill change	Region	No. modes: X, Y, CCA; %Var X, Y	Initial skill	CCA: Skill change
N North America	5, 5, 3 85, 92	0.25	−0.14	South Africa	5, 5, 3 89, 84	0.40	−0.03
S North America	5, 5, 3 86, 83	0.27	−0.12	NW Asia	5, 5, 3 88, 94	0.14	−0.23
South America	5, 5, 3 86, 76	0.37	−0.04	SW Asia	5, 5, 3 89, 83	0.30	−0.01
Greenland	5, 5, 3 86, 95	0.43	0.00	NE Asia	5, 5, 3 80, 85	0.14	−0.10
Europe	5, 5, 3 85, 91	0.18	−0.14	SE Asia	5, 5, 3 83, 87	0.30	−0.07
North Africa	5, 5, 3 86, 89	0.35	−0.03	Indonesia	5, 5, 3 95, 87	0.40	0.01
W Trop Africa	5, 5, 3 87, 88	0.42	0.00	Australia	5, 5, 3 84, 85	0.20	−0.05
E Trop Africa	5, 5, 3 93, 90	0.38	−0.05	Single globe	8, 10, 8 87, 83	0.27	−0.071

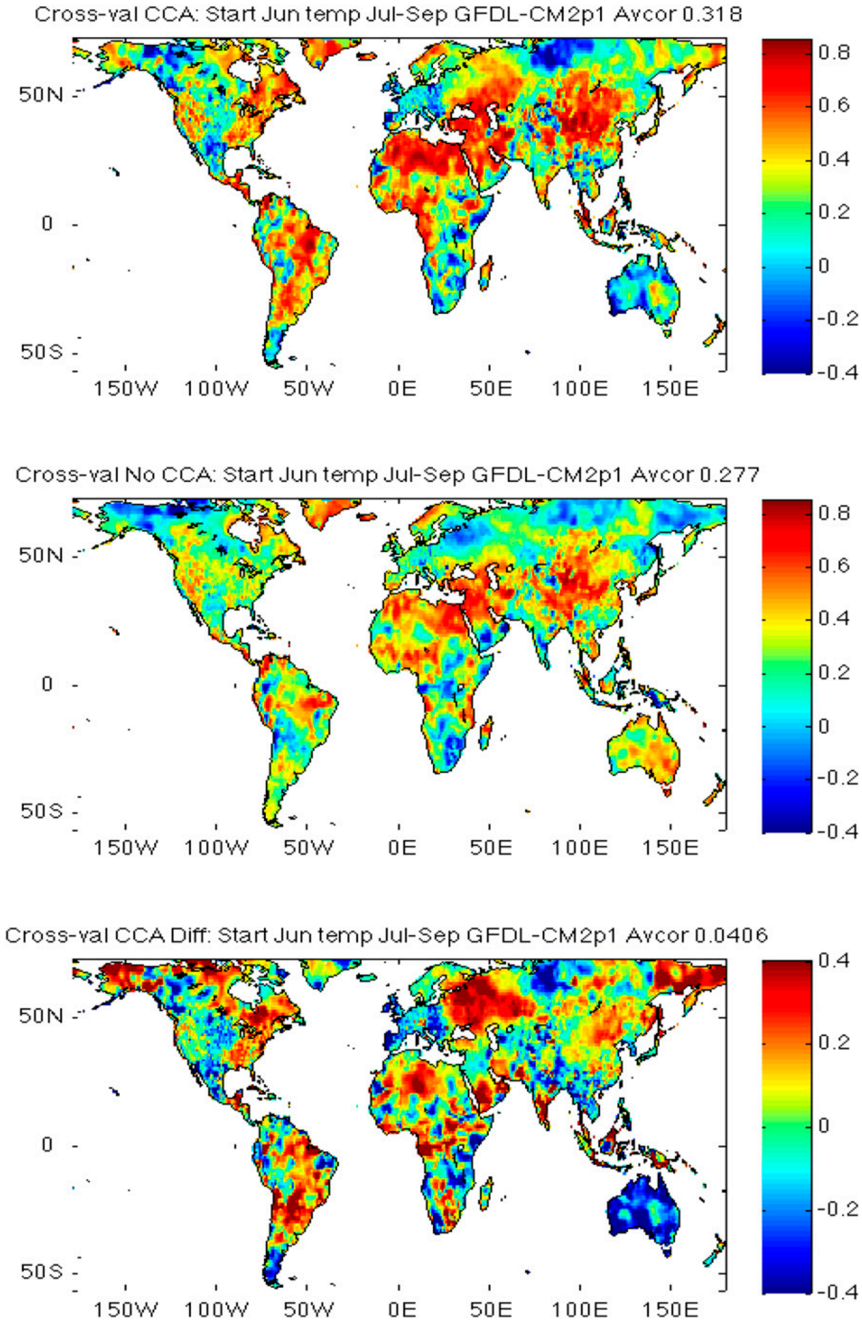

Cross-val CCA: Start Jun temp Jul-Sep GFDL-CM2p1 Avcor 0.318
Cross-val No CCA: Start Jun temp Jul-Sep GFDL-CM2p1 Avcor 0.277
Cross-val CCA Diff: Start Jun temp Jul-Sep GFDL-CM2p1 Avcor 0.0406
50N
0
50S
150W
100W
50W
0E
50E
100E
150E
0.8
0.6
0.4
0.2
0
-0.2
-0.4

FIG. 10. As in Fig. 8, but for temperature forecasts by the GFDL-CM2p1 model for July–September made in early June.

January–March, the single global CCA does not perform better than the individual regional CCAs merged into global forecasts as it does for precipitation. In forecasts of January–March from early October, skills begin slightly lower than those of forecasts starting from early December, and the CCA correction does not result in skill improvements when averaged over all models and regions. Nontrivial skill improvements due to the CCA are only occasional. On the other hand, the skill of forecasts for July–September, which are degraded relatively less by the CCA than those for January–March, shows the single global CCA forecasts outperforming the merged regional CCAs, with skill change near or just above zero.

When RMSESS is used as the verification measure, the CCA strongly improves the skill of the uncorrected temperature forecasts. Figure 11 shows the spatial distribution

TABLE 6. Comparison of the effect on globally averaged anomaly correlation skill of the CCA when performed on individual regions and merged to a global temperature forecast and when performed on the globe as a single region. Results are averaged over all eight models and are shown for forecasts for January–March made in early December and early October and forecasts for July–September made in early June and early April.

Temperature start → target	Original model skill	CCA style	Change from CCA
Dec → JFM	0.273	Merge	−0.070
		Single globe	−0.071
Oct → JFM	0.233	Merge	−0.045
		Single globe	−0.081
Jun → JAS	0.311	Merge	−0.030
		Single globe	0.011
Apr → JAS	0.264	Merge	−0.024
		Single globe	0.000

of RMSESS of uncorrected and CCA-corrected temperature forecasts of the globe when treated as a single region, for the January–March season made from early December. Table 7 shows RMSESS for uncorrected and corrected forecasts for January–March and July–September, each at 1.5- and 3.5-month lead times. In all cases, large improvements in RMSESS are noted. Testing the statistical significance of the difference between uncorrected and corrected RMSESS, using the F ratio of the uncorrected to the corrected $\mathrm{MSE}_{\mathrm{fct}}$ variance terms in Eq. (1), nearly all of the individual grid points are significant at $p < 0.01$, and many at $p < 0.001$, for the four cases shown in Table 7, and field significance for the entire map is even stronger given the multiple spatial degrees of freedom. However, the RMSESS of the corrected forecasts is only slightly positive in all four cases, varying from 0.05 for the longer lead to 0.06 or 0.07 for the shorter lead, showing that the MSE of the corrected forecasts is only slightly smaller than that produced by perpetual climatology forecasts.

The statistical significance of the global average difference from zero of the global average corrected RMSESS for temperature is tested, again using the F test, this time on the variance ratio of the corrected $\mathrm{MSE}_{\mathrm{fct}}$ to $\mathrm{MSE}_{\mathrm{clim}}$. Based on approaches used in Van den Dool and Chervin (1986) and Moron et al. (2007), the spatial degrees of freedom for global temperature are estimated at 15 for January–March and 20 for July–September. The correlation of uncorrected and corrected forecast errors is also computed, representing the lack of independence of the uncorrected and corrected samples because both share the same observations. This second factor works to increase significance when properly accounted for, as shown in DelSole and Tippett (2014). The RMSESS of 0.07 and 0.06 for January–March and July–September, respectively, are found to be significant at the 0.10 level, while the RMSESS of 0.5 for the longer-lead forecasts for the two seasons does not achieve 0.10 level significance. While the short-lead RMSESS significances are weak, they suggest that the positive skill averages for temperature after the CCA correction are unlikely to be positive due only to sampling variability.

Contrasting the RMSESS improvement in temperature forecasts with that noted for precipitation (Fig. 9; Table 4), RMSESS for uncorrected temperature forecasts averages lower than that for precipitation and after correction attains higher levels than for precipitation. The higher final skill might be expected in view of the greater inherent predictability of temperature than precipitation, partly due to its more coherent, less spatially noisy character. Temperature also has the benefit of the largely predictable upward trend due to greenhouse gas increases. Part of the reason for the lower RMSESS for uncorrected temperature than uncorrected precipitation forecasts may be related to the nature of the GHCN-CAMS observed temperature data. Because there is a problem of missing temperature data in much of the developing world (e.g., parts of South America and Africa),[4] the gridding of the GHCN-CAMS data sometimes requires interpolation or extrapolations over long distances. In proportion to the uncertainty in these ‘‘filled in’’ grid squares, the interannual variability may become unrealistically small. In fact, in a few grid squares the interannual standard deviation is found to be zero, and Eq. (1) is not usable. At grid squares having variability smaller than that likely to actually prevail, model biases may be amplified when expressed using Eq. (1), because the RMSE of the perpetual climatology forecasts is unrealistically small when the variability of the observations is low, and the model forecasts are standardized using the standard deviation of the observations.

Considering the nature of handling of missing temperature data in the GHCN-CAMS dataset, we used an alternative temperature dataset to assess skill sensitivity, despite that the anomaly correlation is not directly affected by the interannual variability of the observations.[5] The temperature experiments were repeated using the Climate Prediction Center’s CAMS temperature anomaly data (Ropelewski et al. 1984) instead of the GHCN-CAMS temperature data. The

[4] For precipitation, missing data are less prevalent because of the use of satellite data since 1979 and because in some developing countries in the tropics, precipitation data are archived for agricultural purposes, while temperature data are not archived.

[5] The correlation may be affected by interpolated data because temperatures estimated at such locations likely deviate from the unknown reality in aspects besides their interannual standard deviation.

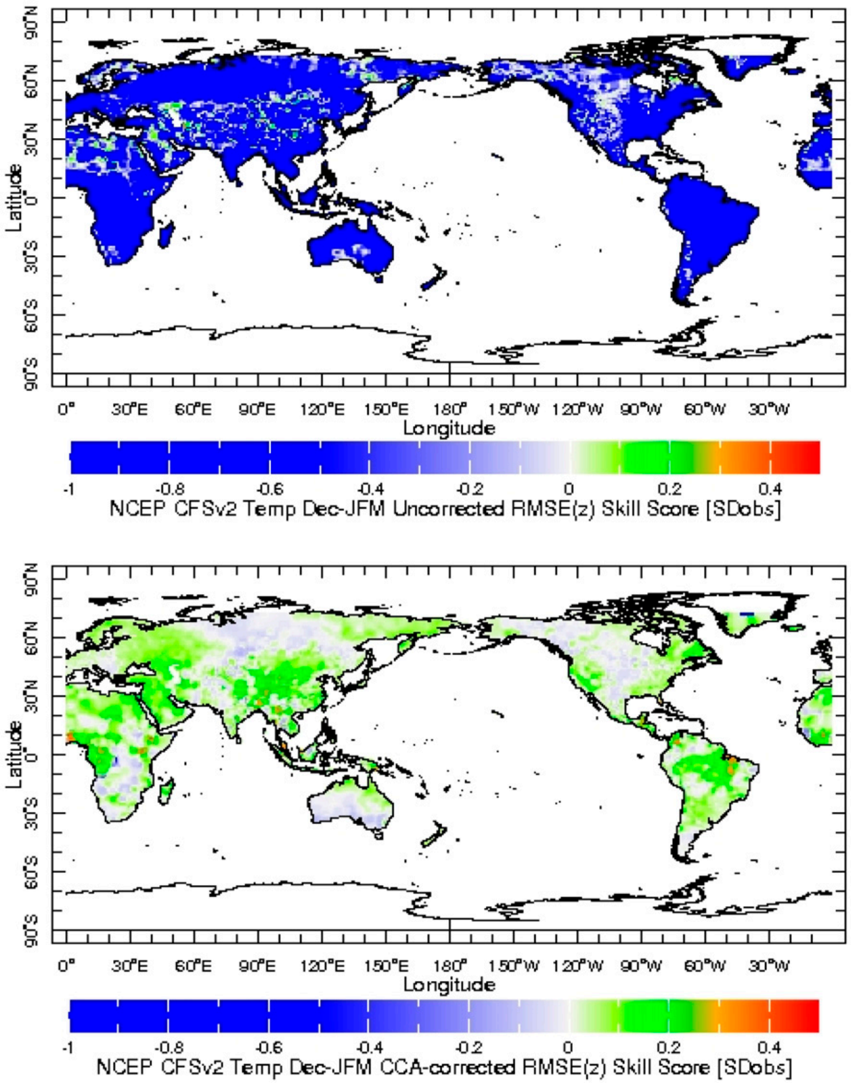

90°N
60°N
30°N
Latitude
0°
30°S
60°S
90°S
0° 30°E 60°E 90°E 120°E 150°E 180° 150°W 120°W 90°W 60°W 30°W
Longitude
-1 -0.8 -0.6 -0.4 -0.2 0 0.2 0.4
NCEP CFSv2 Temp Dec-JFM Uncorrected RMSE(z) Skill Score [SDobs]
90°N
60°N
30°N
Latitude
0°
30°S
60°S
90°S
0° 30°E 60°E 90°E 120°E 150°E 180° 150°W 120°W 90°W 60°W 30°W
Longitude
-1 -0.8 -0.6 -0.4 -0.2 0 0.2 0.4
NCEP CFSv2 Temp Dec-JFM CCA-corrected RMSE(z) Skill Score [SDobs]

FIG. 11. Geographic distribution of RMSESS over the globe as a single region, for temperature forecasts by the NCEP-CFSv2 model for January–March made in early December. (top) The original skill and (bottom) the skill following the CCA correction. The RMSESS is in terms of standardized anomalies with respect to the observed mean and standard deviation.

CAMS anomaly data are on a 2° × 2° grid but were regridded to the 1° × 1° grid for the analyses. Substantial portions of Africa and South America are missing in CAMS, while interpolated or extrapolated in GHCN-CAMS.

Model correlation skill before the CCA correction is higher when using CAMS instead of GHCN-CAMS. This may be due partly to the noise-filtering effect of being based on 2° × 2° grid squares. It may also result from the exclusion of areas of missing data, as opposed to relying on the possibly unrealistic estimations used in the GHCN-CAMS data. The CCA generally improves the correlation skill approximately as much when using CAMS as when using GHCN-CAMS, therefore bringing the skills of most of the regions, and the globe, to higher final skill levels than when using GHCN-CAMS. For example, for the globe as a single region, the initial correlation skill is 0.03 (0.06) higher for CAMS than for GHCN-CAMS in the 1.5-month lead forecasts of January–March (July–September) and higher by 0.05 (0.04) following the CCA. These improved skills suggest that the practical solution of filling in missing temperature data using extrapolation/interpolation over large distances may result in negative biases in globally averaged model skill estimates.

TABLE 7. Global average RMSESS for temperature before and after the CCA correction. Results are shown for forecasts for January–March made in early December and early October and forecasts for July–September made in early June and early April.

Temperature start → target	Global avg RMSESS before CCA	Global avg RMSESS after CCA
Dec → JFM	−3.27	0.07
Oct → JFM	−3.29	0.05
Jun → JAS	−3.14	0.06
Apr → JAS	−3.15	0.05

4. Conclusions and discussion

Using CCA, statistical postprocessing is applied to the hindcasts of the individual models in the NMME, with a focus on the correction of biases in the positions and amplitudes of the predicted large-scale anomaly patterns. The conclusions to follow apply to each of the eight models examined, as in no case do any models show substantial departures from the findings despite differences in details.

The CCA-based corrections are not found to materially improve the anomaly correlation skills of precipitation or temperature forecasts of the individual models of the NMME in the case of most regions, seasons, and lead times. Initial (uncorrected) forecast skills are generally lower for precipitation than for temperature, and improvements are somewhat more favorable for precipitation than temperature. For precipitation, slight improvements are noted for about half of the models for most of the regions and for the globe, for the main target seasons and lead times tested. Most of these skill changes are not statistically significant. Positive outcomes are noted more in cases of models with relatively low uncorrected anomaly correlation, suggesting that the lower skills of these models may be due to systematic pattern errors that are statistically correctable. The effect of the CCA is more substantially positive for short-lead precipitation forecasts for October–December in Indonesia and eastern tropical Africa, where the improvements in some of the models are statistically significant. The rainfalls in both regions are governed substantially by the ENSO state.

Although the temporal anomaly correlation generally is not materially improved by the CCA, the RMSESS is strongly and statistically significantly improved. This result suggests the presence of local biases in the forecasts, such as mean bias and amplitude bias, which degrade the RMSESS but not the anomaly correlation. Pattern placement errors would be expected to degrade both the correlation and the RMSESS. One would expect the CCA to diminish both local systematic biases and spatial placement errors together, even if treatment of both types of errors might require retaining more pre-EOF and CCA modes than if mainly just one type of error were present. The EOF and CCA mode truncation sensitivity tests allow selection of truncations resulting in approximately the best cross-validated correlation skills. Therefore, it can be concluded that the CCA is treating both local calibration biases and spatial placement and amplitude errors and that the latter are either a smaller portion of the total systematic error of the NMME models or require a larger sample size to be better identified. Because of the EOF and CCA mode truncations, the CCA correction is expected to decrease the amplitude of the forecasts, especially given that the NMME forecasts are found generally to have amplitudes higher than that which would minimize squared errors (Barnston et al. 2017) and optimize probabilistic reliability (Van den Dool et al. 2017). This damping tendency alone would serve to weaken the extent of the negative RMSESS results in the uncorrected forecasts. In the case of precipitation forecasts, the large improvement in RMSESS with the CCA correction brings the average RMSESS nearly up to zero, while for temperature a slightly positive (and weakly statistically significantly different from zero) average RMSESS level is attained.

A possible explanation for the lack of skill improvement with CCA-based pattern corrections is that the NMME models are already doing a very good job reproducing nature's large-scale anomaly patterns and their variability. To provide evidence for this possibility, the EOF mode loading patterns for some of the predictions are examined for reasonably mutual comparability. Figure 12 shows the first three modes of precipitation forecasts from the CFSv2 model for January–March for southern North America made in early December and of the corresponding January–March observations. This and other examples indicate that out of the approximately five or six modes retained for most of the regions, only the leading one or sometimes two modes of the model-predicted predictand clearly resemble the corresponding modes of the observed predictand.[6] Thus, the possibility that the models are already reproducing reality quite well is not backed up by a strong correspondence between the EOF modes of model forecasts

[6] Caveats in such a comparison are 1) the observed domain is smaller than the model-predicted domain, 2) the ensemble mean model prediction is more noise-filtered than the single observation, and 3) a poor-to-fair correspondence of modes may be considered benign if one allows for linear combinations of higher-order modes of the forecast set being comparable to linear combinations of such modes in the observed set (a situation that the CCA modes may reveal). A consequence of this third point is that the model's explained variance in terms of the observational EOFs may be high even if the corresponding modes of model and observational EOFs do not look similar.

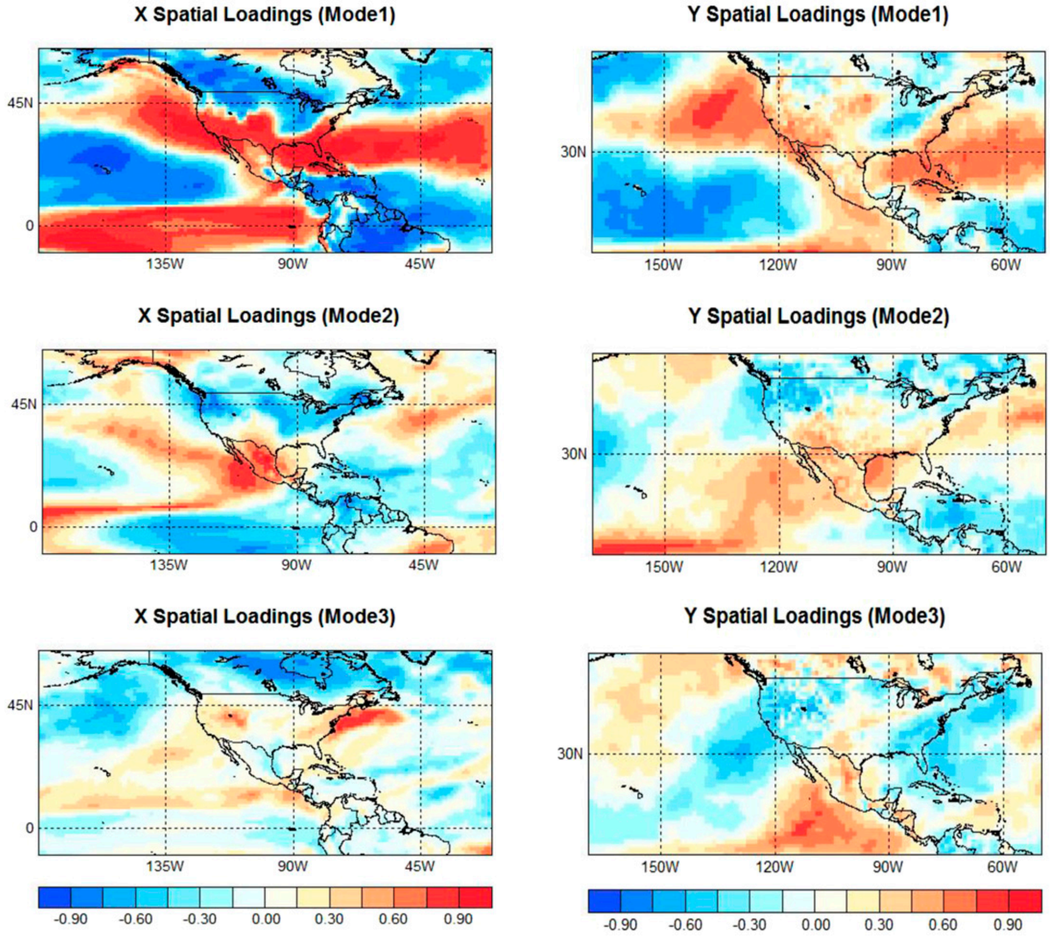

X Spatial Loadings (Mode1)
Y Spatial Loadings (Mode1)
X Spatial Loadings (Mode2)
Y Spatial Loadings (Mode2)
X Spatial Loadings (Mode3)
Y Spatial Loadings (Mode3)
45N
0
30N
135W
90W
45W
150W
120W
90W
60W
-0.90
-0.60
-0.30
0.00
0.30
0.60
0.90

FIG. 12. (left) Spatial patterns for the first three pre-EOFs of the CFSv2 model precipitation forecasts for January–March made in early December for southern North America and (right) the corresponding three pre-EOFs of the observations of the January–March observations. The percentages of original variance explained by the first three modes of the forecasts are 45%, 14%, and 7%, while for the observations they are 21%, 10%, and 9%.

and observations. It is then possible that the models have room for improvement in representing nature's large-scale patterns but that the 29-yr record length used here is insufficient to identify the models' patterns and their corresponding observed patterns robustly enough to perform a beneficial statistical correction. The limited sample issue was suspected to have caused the inability to identify corresponding EOF and SVD mode patterns in 500-hPa heights in model forecasts and observations beyond very short lead times in Rukhovets et al. (1998). The short record greatly limited the statistical significance of many of the results shown here, even in the relatively favorable cases of skill improvement.

While performing the corrections on 15 separate regions and merging them into a global forecast was expected to produce a more skillful corrected global forecast than doing one correction on the globe as a single region, results indicate that the opposite is most often the case for precipitation forecasts and that the two styles produce approximately equal skill results for temperature. It is possible that some of the patterns in the forecasts and observations are sufficiently global that restricting the predictands to subregions prevents some noise filtering that is possible with the global domain and that noise filtering is more critical for precipitation than temperature owing to the smaller signal-to-noise ratio in precipitation. This global scale might help in pattern-based methods like CCA, in contrast to statistical methods that treat one predictand grid square at a time (e.g., multiple regression or principal component regression) even if opting for global predictors.

It is found that covariance-based EOFs lead to better CCA-based skill improvements than correlation-based EOFs for precipitation forecasts, while the opposite is the case for temperature. One might expect the correlations

TABLE 8. Improvement in anomaly correlation due to the CCA for forecasts of January–March precipitation made in early December for various regions and for the globe when using the full, 3-years-out cross validation as used in this study, when using cross validation for CCA model building but using all 29 years to define the pre-EOF modes for X and Y, and when using no cross validation at all.

Precipitation domain	Full cross validation	Cross-validation except for pre-EOFs	No cross validation
Australia	−0.14	−0.12	0.16
Europe	−0.05	−0.03	0.32
Indonesia	0.01	0.07	0.20
S Africa	0.04	0.07	0.31
S America	−0.04	0.06	0.29
S North America	0.01	0.09	0.29
Globe	0.00	0.02	0.57

to do better in most any case, as the strengths of association (implying predictability) are represented directly, regardless of interannual variability. A likely reason that covariances lead to better corrected skills in precipitation is that predictive skill is greatest in the tropics, where ENSO effects are most prominent, and the mean and variance of precipitation are also greatest in the tropics. Covariance-based EOFs thus weight the region of highest skill most heavily. As noted above in the cases of Indonesia and eastern tropical Africa, locations with higher initial skill, particularly when related to a known source (e.g., ENSO), tend to be most amenable to CCA-based skill increases. While temperature predictive skill may also be higher in the tropics than elsewhere, its interannual variability is lower in the tropics than elsewhere, so covariance-based EOFs would work against maximizing average skill.

A factor that might help account for the apparently modest ability of CCA to improve the model anomaly correlation skills for both precipitation and temperature forecasts is the negative skill bias that can occur in cross validation (Barnston and Van den Dool 1993), given the small sample size (29 years) and the large areas of inherently low skill in many of the regions. This bias comes about because, when the correlation over the full sample of years is near zero, and one or more cases are held out from the sample and used as the forecast targets using a prediction model and climatology built from the remaining cases, the relationships in the remaining cases are likely to be of opposite sense to those in the one or more years held out. In other words, there is a complementarity and opposing relationship between the in-sample and out-of-sample cases when the net relationship in the full sample is negligible. Locations with near-zero skill in the total sample can therefore have strongly negative skill after cross validation. Table 8 shows average anomaly correlation changes with the CCA correction, averaged over all eight models for January–March precipitation forecasts made in early December, for various regions and the globe under three cross-validation designs: 1) full cross validation (as used here), 2) cross validation of the CCA regression model but not of the pre-CCA EOFs, which are derived from the full dataset and used for forecasting all years, and 3) no cross validation at all. Skills are higher when the pre-EOFs are not subjected to cross validation while the CCA does use cross validation. This option, while used in some other studies, was declined because in real-time forecasting one does not have the option of forming EOFs that include observed data for the year being forecast. The much higher skill resulting from no cross validation at all illustrates the great extent of overfitting inherent in regression instruments and the necessity of cross validation to get reasonable estimates of skill expected for independent forecast targets.

In conclusion, the general inability of the CCA to systematically improve the correlation skills of any of the individual NMME models, but the strong improvements in the RMSESS measure, mean that the answer to the question posed in the title of this paper is "not much, with just 29 years of data," and suggests the presence of systematic biases largely not on the pattern level. Treatment of such biases may be done by methods less multivariate than CCA, such as principal component regression, multiple regression, or even local simple regression, all of which treat one predictand point at a time.

Acknowledgments. We appreciate the careful and critical reviews by three anonymous reviewers who enhanced the quality of the paper. We are also grateful for the effort of Shuhua Li in initial testing of the CCA software and for the thoughtful insights and suggestions of Simon Mason, Huug van den Dool, Jon Gottschalck, and Dan Collins. The authors were supported by NOAA's Climate Program Office's Modeling, Analysis, Predictions, and Projections program Award NA14OAR4310188. The NMME project and data dissemination is supported by NOAA, NSF, NASA, and DOE. We acknowledge the help of CPC, IRI, and NCAR personnel in creating, updating, and maintaining the NMME archive.

REFERENCES

Barnett, T. P., and R. Preisendorfer, 1987: Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis. *Mon. Wea. Rev.*, **115**, 1825–1850, doi:10.1175/1520-0493(1987)115<1825:OALOMA>2.0.CO;2.

Barnston, A. G., 1994: Linear statistical short-term climate predictive skill in the Northern Hemisphere. *J. Climate*, **7**, 1513–1564, doi:10.1175/1520-0442(1994)007<1513:LSSTCP>2.0.CO;2.

——, and H. M. van den Dool, 1993: A degeneracy in cross-validated skill in regression-based forecasts. *J. Climate*, **6**, 963–977, doi:10.1175/1520-0442(1993)006<0963:ADICVS>2.0.CO;2.

——, M. K. Tippett, M. Ranganathan, and M. L. L'Heureux, 2017: Deterministic skill of ENSO predictions from the North American Multimodel Ensemble. *Climate Dyn.*, doi:10.1007/s00382-017-3603-3, in press.

Benjamini, Y., and Y. Hochberg, 1995: Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Stat. Soc.*, **57B**, 289–300.

Bjerknes, J., 1969: Atmospheric teleconnections from the equatorial Pacific. *J. Phys. Oceanogr.*, **97**, 163–172, doi:10.1175/1520-0493(1969)097<0163:ATFTEP>2.3.CO;2.

Bretherton, C. S., C. Smith, and J. M. Wallace, 1992: An intercomparison of methods for finding coupled patterns in climate data. *J. Climate*, **5**, 541–560, doi:10.1175/1520-0442(1992)005<0541:AIOMFF>2.0.CO;2.

DelSole, T., and M. K. Tippett, 2014: Comparing forecast skill. *Mon. Wea. Rev.*, **142**, 4658–4678, doi:10.1175/MWR-D-14-00045.1.

Delworth, T. L., and Coauthors, 2006: GFDL's CM2 global coupled climate models. Part I: Formulation and simulation characteristics. *J. Climate*, **19**, 643–674, doi:10.1175/JCLI3629.1.

Fan, Y., and H. van den Dool, 2008: A global monthly land surface air temperature analysis for 1948–present. *J. Geophys. Res.*, **113**, D01103, doi: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, doi:10.1175/1520-0442(1999)012<1974:ROMSEB>2.0.CO;2.

Gent, P. R., and Coauthors, 2011: The Community Climate System Model version 4. *J. Climate*, **24**, 4973–4991, doi:10.1175/2011JCLI4083.1.

Goddard, L., and N. E. Graham, 1999: The importance of the Indian Ocean for simulating precipitation anomalies over eastern and southern Africa. *J. Geophys. Res.*, **104**, 19 099–19 116, doi:10.1029/1999JD900326.

Hays, W. L., 1973: *Statistics for the Social Sciences*. Holt, Rinehart and Winston, Inc., 954 pp.

Infanti, J. M., and B. P. Kirtman, 2016: Prediction and predictability of land and atmosphere initialized CCSM4 climate forecasts over North America. *J. Geophys. Res. Atmos.*, **121**, 12 690–12 701, doi:10.1002/2016JD024932.

Johansson, A., A. Barnston, S. Saha, and H. M. van den Dool, 1998: On the level and origin of seasonal forecast skill in northern Europe. *J. Atmos. Sci.*, **55**, 103–127, doi:10.1175/1520-0469(1998)055<0103:OTLAOO>2.0.CO;2.

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

Livezey, R. E., and W. Y. Chen, 1983: Statistical significance and its determination by Monte Carlo techniques. *Mon. Wea. Rev.*, **111**, 46–59, doi:10.1175/1520-0493(1983)111<0046:SFSAID>2.0.CO;2.

Merryfield, W. J., and Coauthors, 2013: The Canadian seasonal to interannual prediction system. Part I: Models and initialization. *Mon. Wea. Rev.*, **141**, 2910–2945, doi:10.1175/MWR-D-12-00216.1.

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

Moron, V., A. W. Robertson, M. N. Ward, and P. Camberlin, 2007: Spatial coherence of tropical rainfall at the regional scale. *J. Climate*, **20**, 5244–5263, doi:10.1175/2007JCLI1623.1.

Murphy, A. H., and E. S. Epstein, 1989: Skill scores and correlation coefficients in model verification. *Mon. Wea. Rev.*, **117**, 572–581, doi:10.1175/1520-0493(1989)117<0572:SSACCI>2.0.CO;2.

Ropelewski, C. F., J. E. Janowiak, and M. S. Halpert, 1984: The Climate Anomaly Monitoring System (CAMS). Climate Analysis Center Tech. Rep., 39 pp.

Rukhovets, L. V., H. M. van den Dool, and A. G. 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, doi:10.1080/07055900.1998.9649606.

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

Smith, T. M., and R. E. 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, doi: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, doi:10.1002/joc.947.

——, T. DelSole, S. Mason, and A. G. Barnston, 2008: Regression-based methods for finding coupled patterns. *J. Climate*, **21**, 4384–4398, doi:10.1175/2008JCLI2150.1.

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

Van den Dool, H. M., and R. M. Chervin, 1986: A comparison of month-to-month persistence of anomalies in a general circulation model and in the Earth's atmosphere. *J. Atmos. Sci.*, **43**, 1454–1466, doi:10.1175/1520-0469(1986)043<1454:ACOMTM>2.0.CO;2.

——, E. Becker, L.-C. Chen, and Q. Zhang, 2017: The probability anomaly correlation and calibration of probabilistic forecasts. *Wea. Forecasting*, **32**, 199–206, doi:10.1175/WAF-D-16-0115.1.

Vecchi, G. A., and Coauthors, 2014: On the seasonal forecasting of regional tropical cyclone activity. *J. Climate*, **27**, 7994–8016, doi:10.1175/JCLI-D-14-00158.1.

Vernieres, G., M. M. Rienecker, R. Kovach, and C. L. Keppenne, 2012: The GEOS-iODAS: Description and evaluation. NASA Tech. Rep. NASA/TM-2012-104606, Vol. 30, 61 pp. [Available online at http://gmao.gsfc.nasa.gov/pubs/docs/Vernieres589.pdf.]

Walker, G., and T. Bliss, 1934: World weather V. *Mem. Roy. Meteor. Soc.*, **4**, 53–84.

Ward, N. 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, doi:10.1175/1520-0442(1997)010<2210:PAOSFV>2.0.CO;2.

Wilks, D. S., 2006: On "field significance" and the false discovery rate. *J. Appl. Meteor. Climatol.*, **45**, 1181–1189, doi:10.1175/JAM2404.1.

——, 2016: "The stippling shows statistically significant grid points": How research results are routinely overstated and overinterpreted, and what to do about it. *Bull. Amer. Meteor. Soc.*, **97**, 2263–2273, doi:10.1175/BAMS-D-15-00267.1.

Xie, P., and P. A. Arkin, 1997: Global precipitation: A 17-year monthly analysis based on gauge observations, satellite estimations, and numerical model outputs. *Bull. Amer. Meteor. Soc.*, **78**, 2539–2558, doi:10.1175/1520-0477(1997)078<2539:GPAYMA>2.0.CO;2.

Zhang, S., M. J. Harrison, A. Rosati, and A. Wittenberg, 2007: System design and evaluation of coupled ensemble data assimilation for global oceanic climate studies. *Mon. Wea. Rev.*, **135**, 3541–3564, doi:10.1175/MWR3466.1.