Final Report for Grant NA14OAR4310188

Improved Probabilistic Forecast Products for the NMME Seasonal Forecast System

Anthony Barnston[1], Michael Tippett[1], Huug van den Dool[2], and Emily Becker[2]

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

[2]Climate Prediction Center, National Centers for Environmental Prediction, National Weather Service, National Oceanic and Atmospheric Administration, College Park, Maryland

Period of Activity: August 1, 2014 to July 31, 2017

Table of Contents

1. Results and Accomplishments

The main purpose of the project is to improve the North American Multimodel Ensemble (NMME) seasonal probabilistic predictions. At IRI, improvements come about from statistical corrections of the prediction patterns of the individual constituent models before combining into the NMME forecasts. At CPC, improvements are derived through a local (non-pattern) calibration of the forecast probability distributions of the models separately at each location. Both aspects require the multidecadal hindcast history of each model, combined with the verifying observations.

CPC Accomplishments

The work at CPC was focused on calibration of probabilities at individual grid locations. In a baseline approach, forecasts were corrected for bias in mean and standard deviation, which improves their skill and probabilistic reliability. Although these calibrations do not account for spatial pattern errors, which would be corrected with IRI's CCA-based correction method, they are a standard that is already hard to beat. The skill and reliability diagnostics for the baseline was summarized in Becker and Van den Dool (2016). A table of probabilistic skill and reliability for surface temperature for the baseline method are shown in the top panel and the top table insert below the two panels in Fig. 1, indicating small but positive average skill, and adequate but not excellent reliability.

The improvements are expected due to local calibrations of the probability forecasts consists largely of managing the histograms (the probability forecast distributions) shown on the right side of Fig. 1. For example, too-bold forecasts are damped when the probabilistic skill does not warrant the degree of deviation from the probability of climatology (0.333).

The method for calibrating the local probabilities is a damping of probability anomalies as per a new verification measure called the probability anomaly correlation, or PAC (Van den Dool et al. 2017). The calibration is found to result in improvements in the Brier skill score and in probabilistic reliability analysis.

A demonstration of the improvement in practice by the probability calibration was achieved using the seasonal forecasts of individual NMME models and of the NMME. Figure 1 shows reliability and forecast frequency histograms for short-lead forecasts of SST in the Northern Hemisphere by the CFSv2 model before (top) and after (bottom) the calibration of the probability anomalies. The calibration markedly improves the reliability for each of the tercile-based categories, as the lines come closer to the ideal y=x line. The frequency histograms show that the improvement is accomplished by damping the probability anomalies—i.e., making the forecasts less sharp, but not unduly so. The Brier skill score is also improved following the calibration (not shown).

Figure 2 shows a similar reliability and forecast frequency analysis, but for all seasons and lead times combined, for the baseline for the NMME, for surface air temperature. Here the probability anomaly calibration has not been applied. Because the inherent predictability is much lower than it is for SST predictions, the sharpness of the forecasts is necessarily modest. The probabilistic reliability of the baseline is not perfect, but adequate, and it would be expected to improve following the calibration. This is indeed found to be the case (not shown). Fig. 2 is important in the sense that expectations should be realistic—i.e., f the attributes diagram looks fairly good already, the PAC based adjustment cannot be expected to yield a much better result.

Figure 3 shows ranked probability skill score (RPSS) maps for NMME lead-1 seasonal probability forecasts of 2 meter temperature over North America from the 1982-2010 hincasts, uncalibrated (top) and PAC-calibrated (bottom). Areas of

negative RPSS (white) are removed, and areas with positive skill scores are slightly increased. RPSS maps for both uncalibrated and calibrated forecasts are created and posted alongside the realtime forecasts, for evaluation by the operational seasonal forecasters.

Discussion with operational forecasters led to two adjustments in the final PAC-based products. First, the forecasters desired forecasts over the oceans, not just over land. The solution was to use the Reynolds optimum interpolation (OI) SST data to represent the air temperature over oceans, to complement the surface air temperature over land. The observations then extended globally, allowing for training of the PAC on the global domain, and resulting in global calibrated forecasts. The second detail was with regard to the damping effect of the PAC calibration, resulting in some loss of forecast coverage, with the weakest forecast probability contour at 40% (as opposed to the "non-forecast" of 33%). The solution, at forecaster request, was to change the lowest contour shown on the new maps to 36%. Following the two adjustments, forecasters were happy to regularly use the improved PAC-calibrated probability forecasts from the NMME.

IRI Accomplishments

At IRI, software was developed to correct individual coupled ocean-atmosphere models for systematic error in the spatial location and amplitude of large scale anomaly patterns, so that patterns in the uncorrected forecast could be modified in shape, strength and position simultaneously to better reflect the most likely corresponding observations. Canonical correlation analysis (CCA) was used for the corrections. Corrections are done on a regional basis, and then later concatenated to form a globally corrected forecast. Originally, 10 regions had been defined, but tests indicated that smaller regions result in more effective correction results, so 15 regions were defined, as shown in Fig. 1. The regions vary in size in accordance with the size of the areas having common sets of responses to global climate signals, such as ENSO. Although such regions would ideally be varied seasonally, a best compromise was reached, weighting the seasons having greatest predictive potential most heavily. The 15 regions overlap to form transition zones having linearly varying weighting of regional inputs, resulting in a smooth final global map. Each model is processed individually, and the corrected forecasts of the models are then averaged into a multi-model forecast.

Figure 2 shows the average correlation between the model precipitation forecasts and the observations before the CCA correction (left side), for forecasts of precipitation in January-March (top two rows) and July-September (bottom two rows). Results for two lead times are shown—one short lead (from early in the previous month), and one longer lead (from two months earlier than the short lead). In each panel, the bars show the results for each of the 8 models. The right side shows the change in the correlation following the CCA correction. Results show small improvements for some models for the winter forecasts, especially at short lead (top right panel), but improvements are seen to be generally modest. Individual examples of the spatial distribution of the correlation skill of the original model

precipitation forecasts, the forecasts after the CCA correction, and the change due to the CCA (not shown) indicate skill improvements in various subregions (or "pockets") across the map, balanced by other subregions in which skill is degraded or left approximately unchanged.

Table 1 shows the initial skill, averaged over all 8 models, and the change in skill associated with the CCA for each of the 15 regions, for winter precipitation forecasts made in early December. The changes due to the CCA correction vary by region, and average -0.02, showing an overall lack of improvement.

A CCA correction for the entire globe as a single region, for the same season and lead time, results in a skill change of 0.000 – better than that for the merging of the 15 regions. This was an unexpected finding, as intuition would suggest that when attention is focused on individual regions, including use of varying predictor regions and numbers of modes used in the CCA, results might be better.

Table 2 shows a comparison of skills for precipitation forecasts for January-March and for July-September, each made at two lead times, first for the 15 regions treated individually and merged to cover the globe, and then for the globe treated as a single region. In all four cases, the single globe correction results in relatively better skill. Therefore, the single globe strategy was favored for most of the remainder of the project. Figure 3 shows the spatial distribution of correlation skill of the original model precipitation forecasts, the forecasts after the CCA correction, and the change due to the CCA, for the globe for the CMC1-CanCM3 model for early December forecasts of January-March. Overall improvement is small (mean improvement is 0.029), but improvements are substantial in some subregions, such as the southwest and northeast US.

While not shown, the CCA results in sizeable improvements in precipitation forecast skill in some regions having their rainy seasons during periods outside of northern hemisphere winter and summer. For example, useful improvements are realized in eastern tropical Africa and for Indonesia for forecasts for October-December made in early September. In both cases, predictability is known to exist in association with the ENSO state, and the CCA helps refine the positions and strengths of the precipitation response patterns across the region.

Although the correlation was intended to be used as the main verification measure for the experiments, the root-mean-squared error skill score (RMSESS) was also used, to see if the CCA correction reduces local systematic errors even if errors in the placement and amplitudes of spatial patterns are not appreciably reduced. Figure 4 shows the RMSESS before and after the CCA correction, and shows a very significant improvement following the CCA. This suggests that the CCA is reducing errors that vary over small spatial scales—maybe even adjacent individual grid points—rather than varying over larger distances as expected with errors in the placement and amplitude of the coherent, large scale anomaly patterns. This unexpected result shows that the CCA is useful, but not in the anticipated manner. Table 3 shows the global average of the RMSESS before and after the CCA for

forecasts of two seasons made at each of two lead times. Dramatic improvements due to the CCA are noted for all four combinations of season and lead time.

The experiment was also applied to temperature forecasts. Temperature is better forecast than precipitation, so skills before the CCA are at a higher starting point. Table 4 shows the starting skill and the skill change due to the CCA for each of the 15 regions (and the globe as a single region), averaged over the 8 models, for January-March temperature forecasts made in early December. The CCA helps the skills for temperature forecast less than it does for precipitation, and only one or two of the 15 regions have a skill improvement due to the CCA. The average of the skill change over the 15 regions, and also for the globe treated as a single region, is -0.07, showing that in an overall sense the CCA is not helpful for temperature forecasts for this season and lead time. Table 5 shows results for winter and summer temperature forecasts at each of two lead times, for a merging of the individually corrected regions and for the globe treated as a single region. Results are unfavorable for forecasts of January-March (as seen in Table 4), and less negative for July-September season, especially for the globe treated as a single region. Even in the best case—forecasts for July-September made in early June—the improvement is modest.

The effect of the CCA on temperature forecasts is very different for the RMSESS (Table 6), where a dramatic improvement in skill comes about due to the CCA. Apparently, as also found above in the case of precipitation, the CCA is reducing errors that vary over small spatial scales rather than errors associated with placement or amplitude errors in the large scale anomaly patterns. The RMSESS improvement due to the CCA, and the final skill, is greater for temperature than for precipitation. Figure 5 shows the geographical distribution of RMSESS for temperature forecasts for January-March made in early December by the NCEP-CFSv2 model before and after the CCA. Across much of the globe, initial systematic errors are very large before the CCA, and dramatically reduced after it.

Examination of the correction of temperature forecasts was extended by using alternative temperature data sets—namely CAMS instead of GHCN-CAMS. Results using CAMS were relatively more favorable, but the CAMS data is inconvenient for operational use because of its large areas of missing data in various regions.

A journal publication describing the results presented herein and in the final year's annual report was submitted to the Journal of Climate (Barnston and Tippett 2017).

The software used for the CCA-based forecast corrections has been sent to CPC, in case the researchers or forecasters are interested in probing further into its utility in reducing systematic local forecast error either using the CCA or a simpler method such as principal component regression or simple regression.

2. Highlights of Accomplishments

- At NOAA/CPC, the probabilistic calibration method was tested, and found to result in substantial improvements in probabilistic forecast verification measures, including reliability. Therefore, the forecasts using the method have been deployed in CPC's realtime NMME forecasts.
- At IRI, systematic errors in individual coupled model forecast spatial patterns were corrected, resulting in skill improvements in specific regions and seasons, but modest results overall. However, improvements in an error score at a local (not pattern) level are substantial for both precipitation and temperature. The CCA was therefore found to be useful for an unintended purpose, as local corrections can be done using simpler methods than CCA. The software used for the CCA-based forecast corrections have been sent to CPC, should they be interested in using it or doing further work with it.

3. Publications and Reports

3.1 Publications by Principal Investigators

Barnston, A. G., and M. K. Tippett, 2017: Do statistical pattern corrections improve seasonal climate predictions in the North American Multi-model Ensemble models? *J. Climate*, **30**, in press.

Becker, E. J., and H. van den Dool, 2016: Probabilistic seasonal forecasts in the North American Multimodel Ensemble: a baseline skill assessment. *J. Climate*, **29**, 3015-3026.

Van den Dool, H., E. Becker, L.-C. Chen and Q. Zhang, 2017: The probability anomaly correlation and calibration of probabilistic forecasts. *Weather and Forecasting*, **32**, 199-206.

3.2 Other Relevant Publications

3.3 Other Relevant Older Publications

4. PI Contact Information

Anthony Barnston, tonyb@iri.columbia.edu 845-680-4447, International Research Institute for Climate and Society, Columbia University, Palisades, New York 10964.

5. Slides

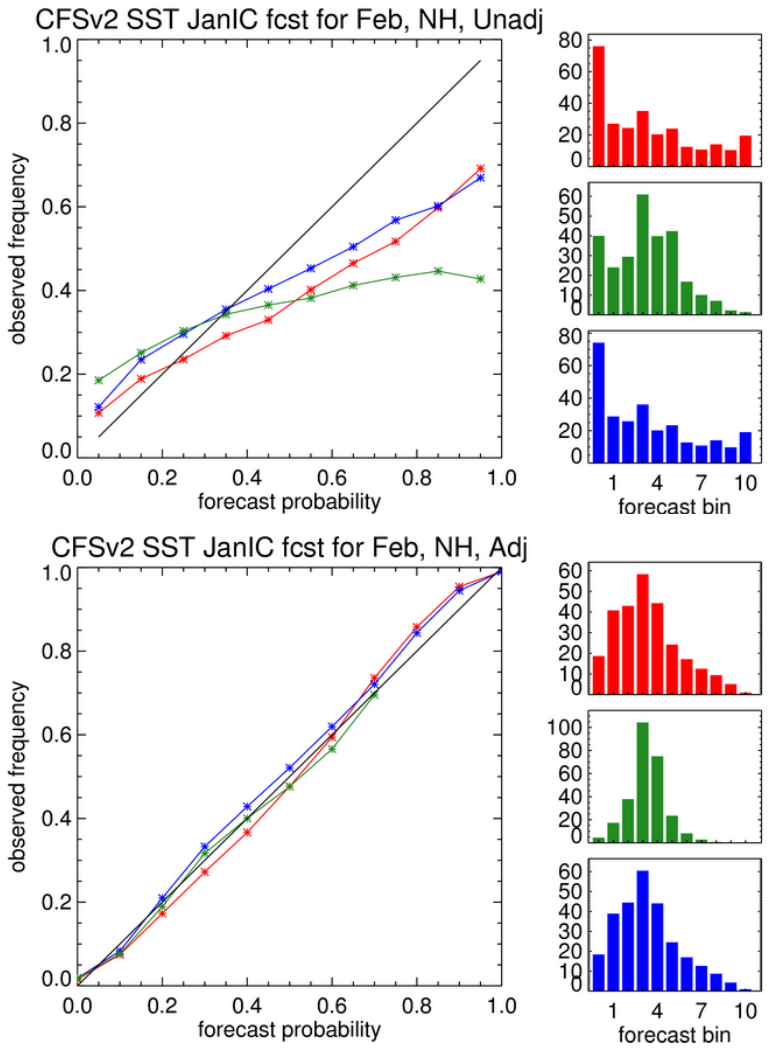

Figure 1. Reliability diagrams for the baseline (top) and adjusted (bottom) NMME hindcasts for Northern Hemisphere extratropics (23N-75N), for each tercile category. The x-axis shows forecast probability, and the y-axis is observed relative frequency. The black line shows ideal reliability, i.e., y=x. On right side are shown histograms of the frequency with which various forecast probabilities are issued for

each of the three categories, showing forecast sharpness. each bar is 0.1.

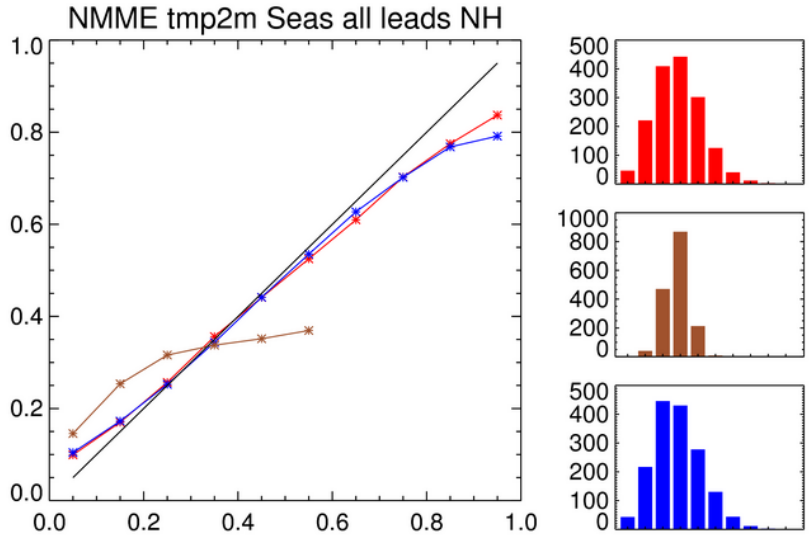

Figure 2. Top: Probabilistic skill score (Brier skill score) for baseline NMME hindcasts for 3-month average surface temperature for all seasons, spanning 1982-2010 for the Northern Hemisphere extratropics (23N-75N), by tercile category and forecast lead time. Middle: Reliability diagram for the baseline NMME surface temperature hindcasts. See caption of Figure 1 for details about the diagram.

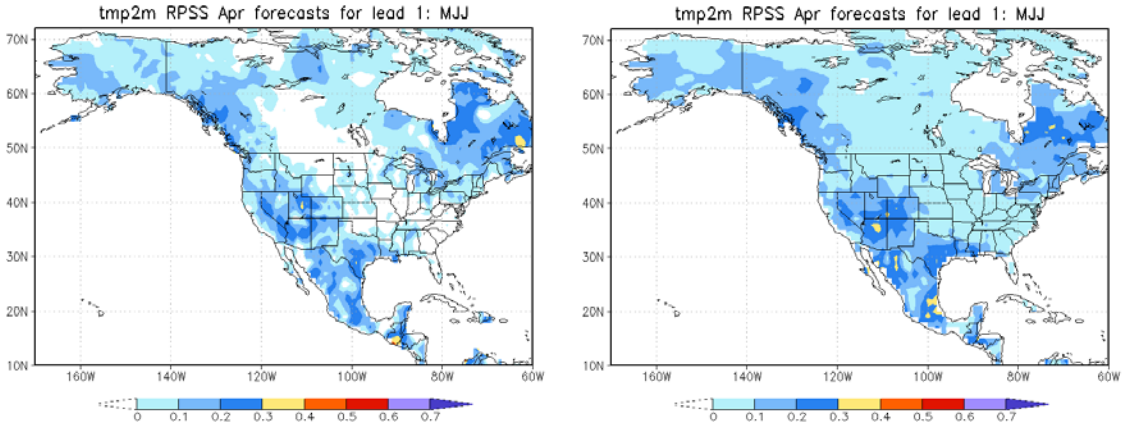

Figure 3. Ranked probability skill score (RPSS) for uncalibrated (upper) and calibrated (lower) NMME probabilistic forecasts of 2 m temperature, lead-1 (MJJ) from April. Calibration is performed through the PAC method. RPSS is calculated based on the 1982-2010 hindcasts.

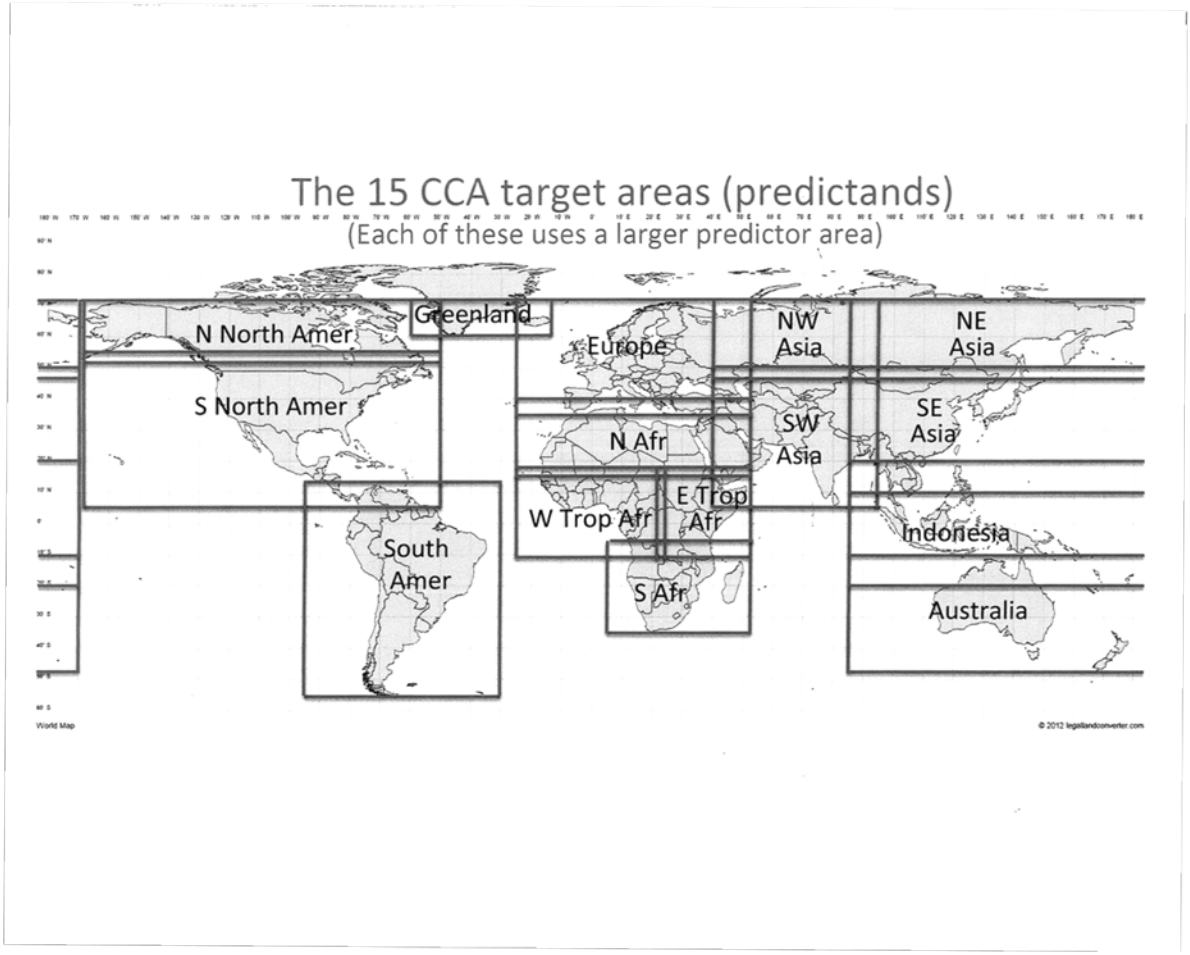

Figure 1. The 15 regions used for the CCA-based model forecast corrections. Overlap regions use forecasts from more than one region, which are averaged using linear location-based weighting.

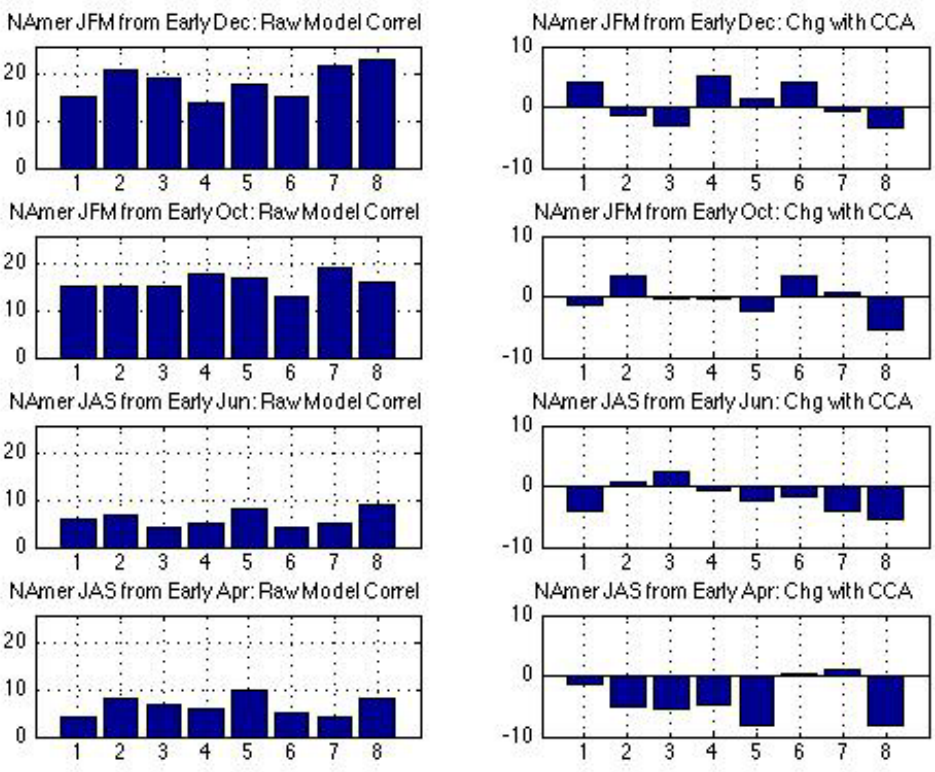

Figure 2. Original anomaly correlation skill X 100 (left), and the change in skill due to the CCA (right) for the non-northern North America region for each of the eight NMME models for precipitation. From top to bottom, the results are for (row 1) January-March precipitation forecasts from early December, (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 8 models (horizontal axis) is (1:CCSM4, 2:NASA, 3:GFDL, 4:GFDL-FLORA, 5:GFDL-FLORB, 6:CMC1, 7:CMC2, and 8:CFSv2).

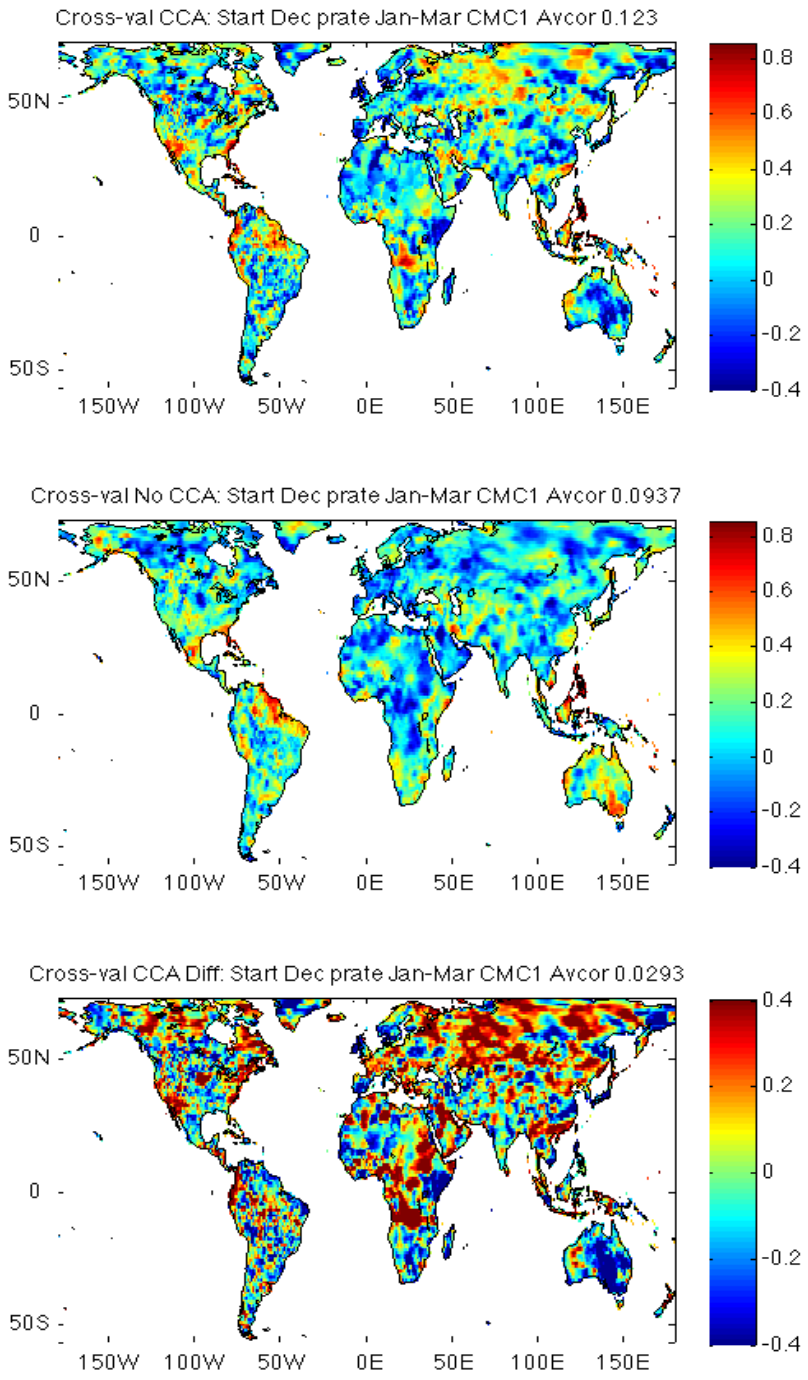

Figure 3. Geographic distribution of temporal anomaly correlation skill over the globe as a single region, for precipitation forecasts by the CMC1-CanCM3 model for January-March made in early December. The middle panel shows the original skill, top panel the skill after the CCA correction, and bottom panel the skill improvement due to the CCA (note the different scale for the bottom panel).

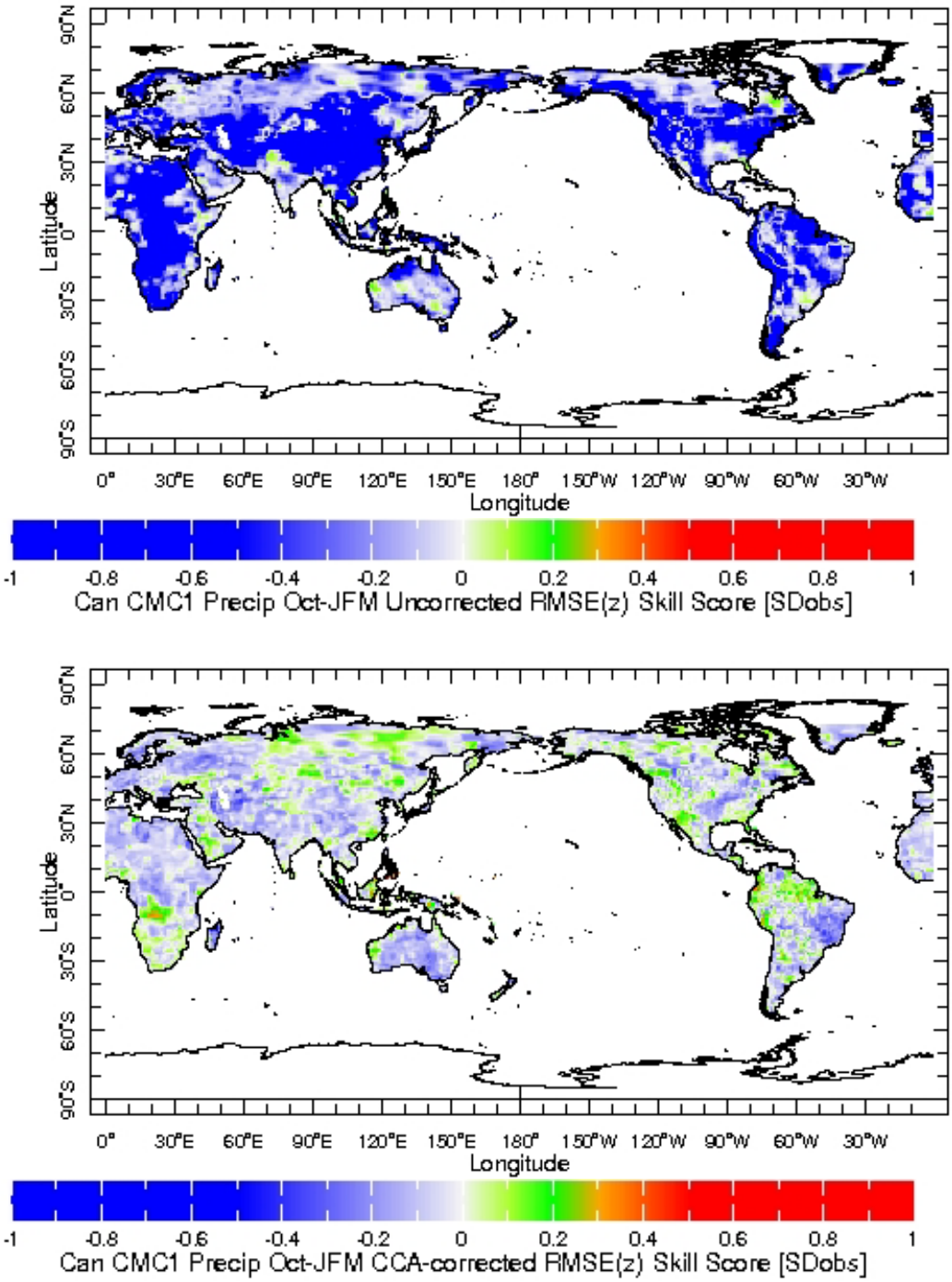

Figure 4. Geographic distribution of root mean squared error skill score (RMSESS) over the globe as a single region, for precipitation forecasts by the CMC1-CanCM3 model for January-March made in early October. The top panel shows the original skill, and the bottom panel the skill following the CCA correction. The RMSESS is in terms of standardized anomalies with respect to the observed mean and standard deviation.

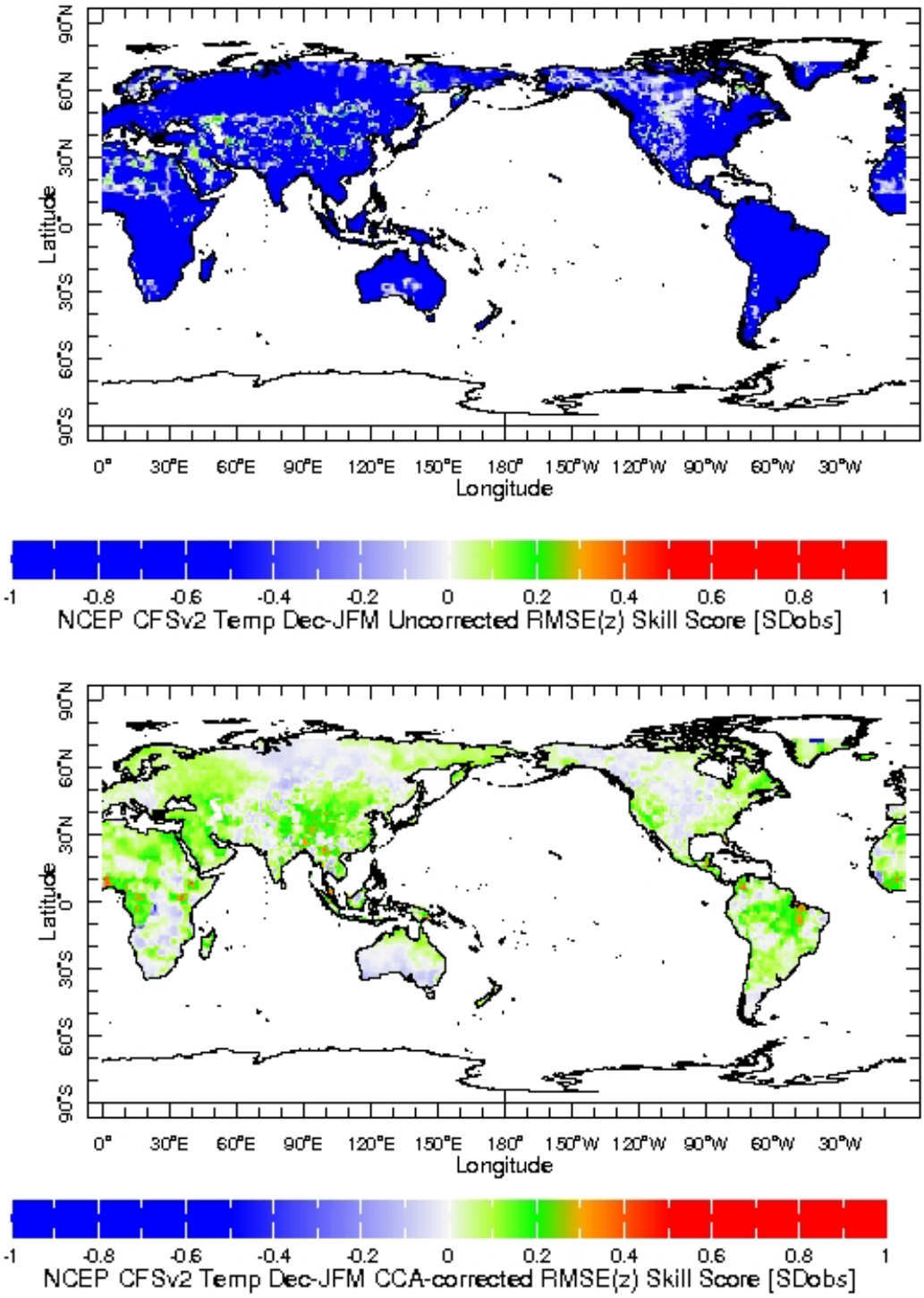

Figure 5. Geographic distribution of root mean squared error skill score (RMSESS) over the globe as a single region, for temperature forecasts by the NCEP-CFSv2 model for January-March made in early December. The top panel shows the original skill, and the bottom panel the skill following the CCA correction. The RMSESS is in terms of standardized anomalies with respect to the observed mean and standard deviation.

Precipitation Region	Initial Skill	CCA: Skill Change	Precipitation Region	Initial Skill	CCA: Skill Change
N North Amer	0.05	0.03	South Africa	0.10	0.04
S North Amer	0.18	0.01	NW Asia	0.10	0.06
South Amer	0.15	-0.04	SW Asia	0.13	-0.06
Greenland	0.06	0.05	NE Asia	0.09	-0.01
Europe	0.07	-0.05	SE Asia	0.12	-0.06
North Africa	0.07	-0.04	Indonesia	0.24	0.01
W Trop Africa	0.01	-0.01	Australia	0.24	-0.14
E Trop Africa	0.05	-0.16	**Single Globe**	**0.114**	**0.000**

Table 1. 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 8 models, for each of 15 individual regions and for the globe as a single region. The area-weighted average change in skill of the 15 individual regions is -0.02.

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

Table 2. 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 8 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	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

Table 3. 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.

Temperature Region	Initial Skill	CCA: Skill Change	Temperature Region	Initial Skill	CCA: Skill Change
N North Amer	0.25	-0.14	South Africa	0.40	-0.03
S North Amer	0.27	-0.12	NW Asia	0.14	-0.23
South Amer	0.37	-0.04	SW Asia	0.30	-0.01
Greenland	0.43	0.00	NE Asia	0.14	-0.10
Europe	0.18	-0.14	SE Asia	0.30	-0.07
North Africa	0.35	-0.03	Indonesia	0.40	0.01
W Trop Africa	0.42	0.00	Australia	0.20	-0.05
E Trop Africa	0.38	-0.05	**Single Globe**	**0.27**	**-0.071**

Table 4. 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 area-weighted average change in skill of the 15 individual regions is -0.07.

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

Table 5. 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 8 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	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

Table 6. 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.