Science and Technology Infusion Climate Bulletin
NOAA’s National Weather Service
38th NOAA Annual Climate Diagnostics and Prediction Workshop
College Park, MD, 21-24 October 2013

Sensitivity Study of the Skill of the CPC Week-2 Reforecast Tool to Reforecast Sampling

Melissa Ou, Mike Charles, Dan Collins, and Emily Riddle

Climate Prediction Center, NCEP/NWS/NOAA, MD

1. Introduction

This study assess the impact of reducing the number of years, number of ensemble members, and frequency of reforecasts on the skill of week-2 calibrated surface temperature and precipitation forecasts using the current Global Ensemble Forecast System (GEFS) model reforecasts. These week-2 forecasts are referred to as the week-2 ‘Reforecast Tool’ and are evaluated over the Contiguous United States (CONUS) for this assessment.

Previously, the NOAA Earth System Research Laboratory (ESRL) had been producing reforecasts, which the Climate Prediction Center (CPC) has used to create 6-10 and 8-14 days calibrated (week-2) forecasts. However, ESRL will no longer be generating these reforecasts, and it was proposed to be created by the Environmental Modeling Center (EMC) at the National Centers for Environmental Prediction (NCEP) therefore enabling reforecasts to be updated more frequently with the real-time GEFS. The EMC requested input from the CPC regarding what they deemed necessary (for CPC’s forecast timescales) in terms of the amount of reforecasts that would be produced in the future by EMC.

To evaluate the sensitivity of week-2 reforecast tool skill to reforecast sampling, 11 configurations (cases) of reforecast sampling are selected to produce skill scores of calibrated week-2 forecasts. The goal is to determine how the skill of the week-2 reforecast tool varies based on various reforecast configurations to help determine what would be considered a sufficient number of reforecasts needed without a significant loss in skill. A lower configuration of reforecasts would require less computational resources on behalf of EMC to produce these reforecasts.

TABLE 1 Explanation of cases used in the study.

Case #	# Training Years	# Ensemble Members	Model Run Frequency
0	1985-2010 (26 yrs)	11	daily
1	2001-2010 (10 yrs)	11	daily
2	1985-2010 (26 yrs)	6	daily
3	1985-2010 (26 yrs)	11	once every 3-4 days (2/week, Mon and Thurs)
4	1985-2010 (26 yrs)	11	once every 7 days (every Thurs)
5	1985-2010 (26 yrs)	6	once every 7 days (every Thurs)
6	2001-2010 (10 yrs)	11	once every 7 days (every Thurs)
7	2001-2010 (10 yrs)	6	once every 7 days (every Thurs)
8	1993-2010 (18 yrs)	6	once every 7 days (every Thurs)
9	1985-2010 (26 yrs)	3	once every 7 days (every Thurs)
10	2001-2010 (10 yrs)	3	once every 7 days (every Thurs)
11	1985-2010 (26 yrs)	1	once every 7 days (every Thurs)

Correspondence to: Melissa Ou, Climate Prediction Center, NCEP/NWS/NOAA, 5830 University Research Court, College Park, MD 20740-3818; E-mail: Melissa.Ou@noaa.gov.

Three different parameters are evaluated with varying configurations of reforecasts to calculate the statistics used to “train” the reforecast tool – the number of training years, the number of ensemble members, and the model run frequency. The model run frequency is the number of times a week a reforecast dataset is used in the statistics calculation for training the reforecast tool.

2. Data and methodology

The reforecasts used in this study are from the Global Ensemble Forecast System (GEFS) with physics operational during 2012, provided by ESRL. The reforecast dataset includes daily reforecasts for 26 years (from 1985-2010) with 11 ensemble members (including a control run). 16 months of real-time GEFS data (from Feb 26, 2012 to June 11, 2013), with physics also operational in 2012 are used in the calibration. Forecasts are formatted as probabilities of three different categories, below-normal, above-normal, and near-normal. The observations used to calculate the skill scores are station-based 5 and 7 day means of CPC’s U.S. station-based daily precipitation and temperature data. About 200 stations of data are utilized.

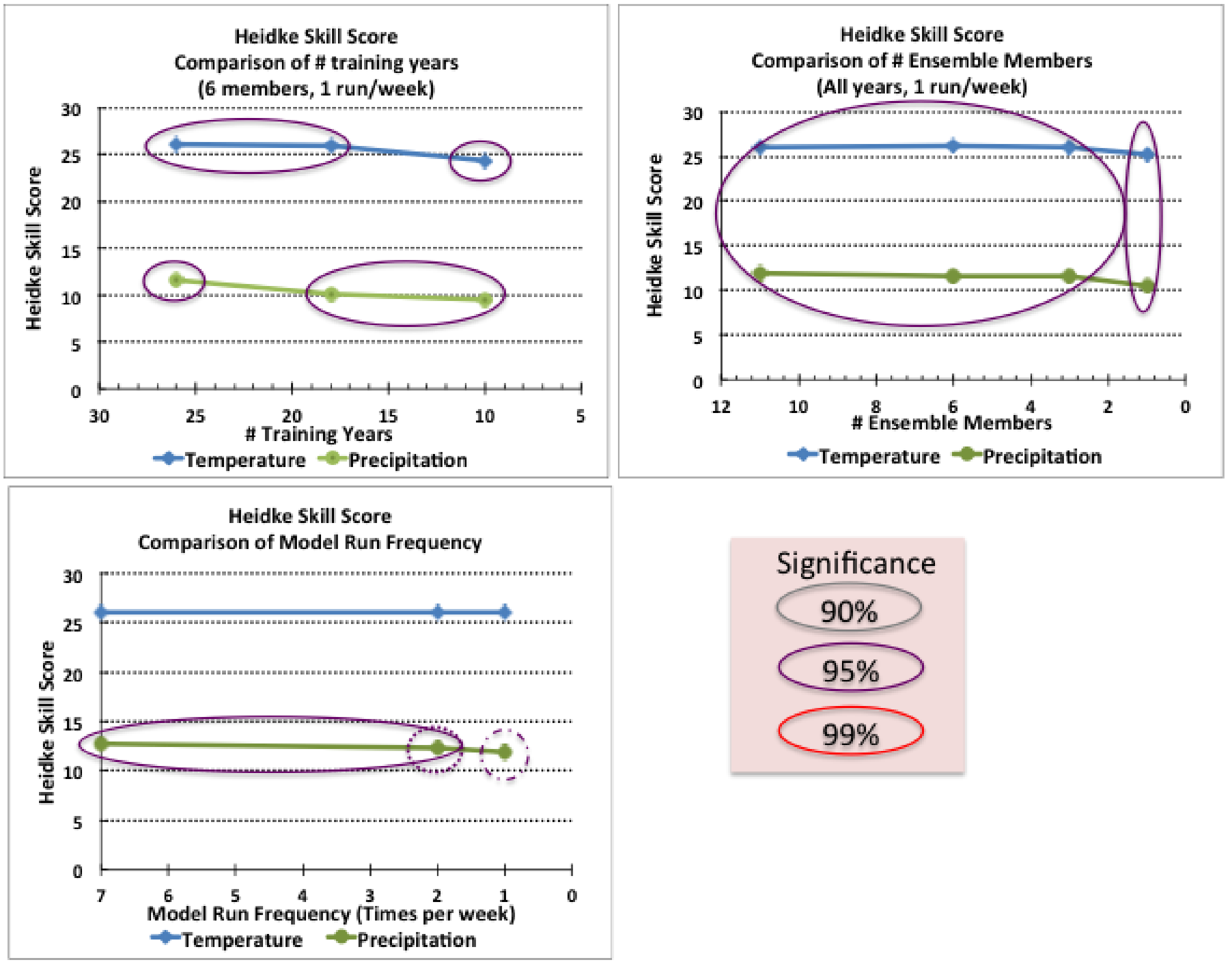

Fig. 1 Line plots of Heidke Skill Score for various configurations of reforecast sampling for 3 different parameters for temperature and precipitation. Pairs of skill scores that have a value difference significant to the 90% level or greater are circled in matching colors. Multiple points in one circle on a plot with a matching circle of point(s) of the same color denotes the skill value difference at significance levels >= 90% for the same variable. The pink box indicates the colors of circles and the associated significance level. For example, in the plot for the # training years (top left), the difference in skill of both 26 and 18 years compared to the skill value of 10 years have a significance >= 95%.

To create forecasts, for each of the cases, first statistics are generated by using the reforecasts and associated observations. The model analysis fields from the reforecasts are used as the “observations” in the calibration to train the model. It should be noted that using an actual observation dataset did not greatly alter verification results, so the model analysis is used for the statistics calculation.

Next, these statistics are used to calibrate the real-time ensemble forecasts (2012-2013) using the ensemble linear regression method (Unger, 2009). This produces tercile probabilistic forecasts of temperature and precipitation. The three categories of the forecast are above-normal, below-normal, and near-normal. Finally, skill scores are generated using CPC's verification system. These steps are done for each of the cases, by sampling the reforecast dataset according to the configuration specified by each case to calculate the statistics that go into the calibration step.

This study is designed so that there are a reasonable number of cases to evaluate the skill to capture a sufficient range of reforecast sample configurations, while keeping the number of cases to a minimum due to the significant computational time it would take to step through the process to create skill scores. 11 cases were created with differing combinations of the three parameters being evaluated. Case 0 represents the maximum configuration, with all 26 years, 11 ensemble members, and daily reforecasts used per week. Subsequent cases use sub-samples of the pool of available reforecasts. Table 1 shows the details of each of these configurations.

Three skill scores are used in this evaluation, including the Heidke Skill Score (HSS), Rank Probability Skill score (RPSS), and the Reliability Skill Score. The HSS and RPSS are calculated by aggregating over the CONUS for each time step of the 16 months of available forecasts, then the mean skill from these time series of scores are computed. A 1-tail two-sample t-test was performed for testing the significance of the mean skill differences from each of the different cases.

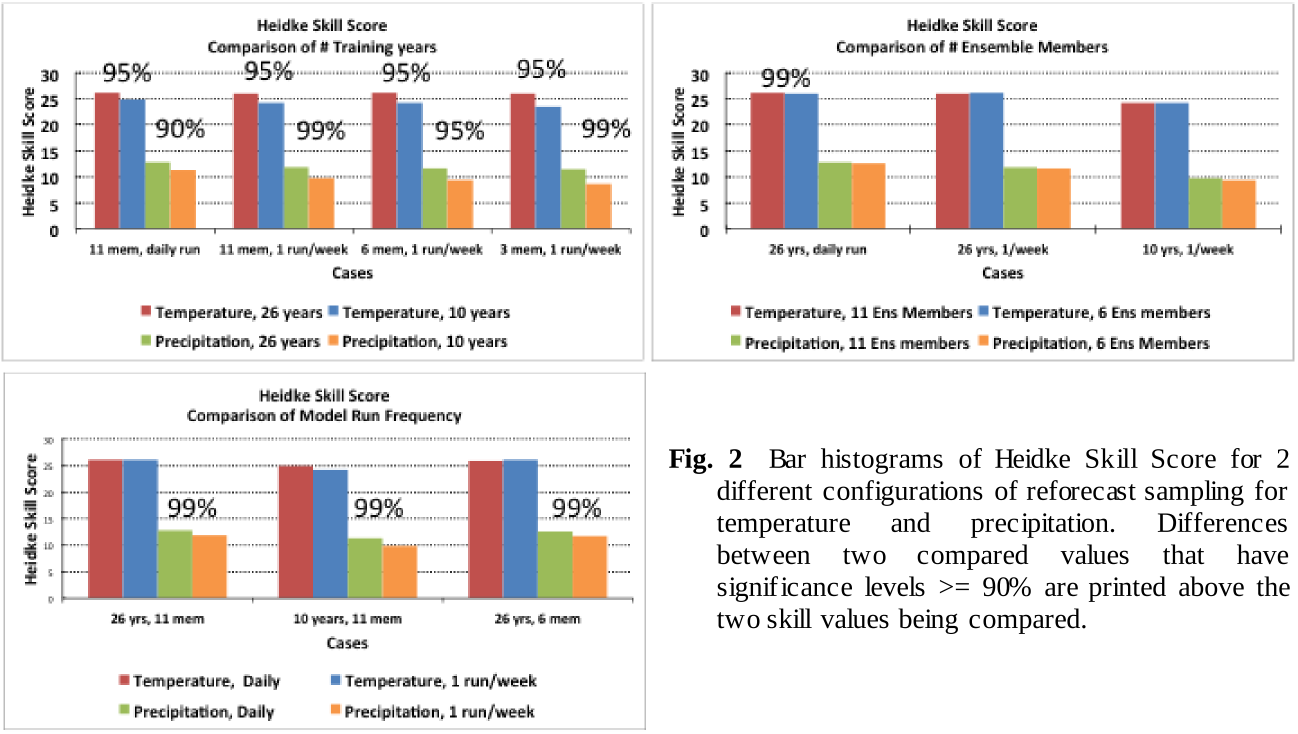

Fig. 2 Bar histograms of Heidke Skill Score for 2 different configurations of reforecast sampling for temperature and precipitation. Differences between two compared values that have significance levels >= 90% are printed above the two skill values being compared.

3. Results

The cases selected for each chart are based on ease of comparing cases isolating impacts by changing each of the parameters. Line plots and reliability diagrams in this study shows how skill of forecasts change with decreasing configurations of each of the three parameters, while keeping the other 2 parameters constant at a selected case. The caveat of using these types of charts is that the skill associated with the decreasing configurations of the changed parameter only reflect one case of the other 2 fixed parameters (*e.g*. changing training years for 26, 18, and 10 years, while keeping the members and model runs constant at 6 and 1, respectively). Bar histograms are also created for Heidke and RPSS, which allows multiple cases of the 2 fixed parameters to be compared but for only 2 different configurations of the changing parameter (*e.g.* for changing the number of training years, 26 *vs*. 10 years are compared for 11 members/daily model run, 11 members/1 run/week, *etc*.).

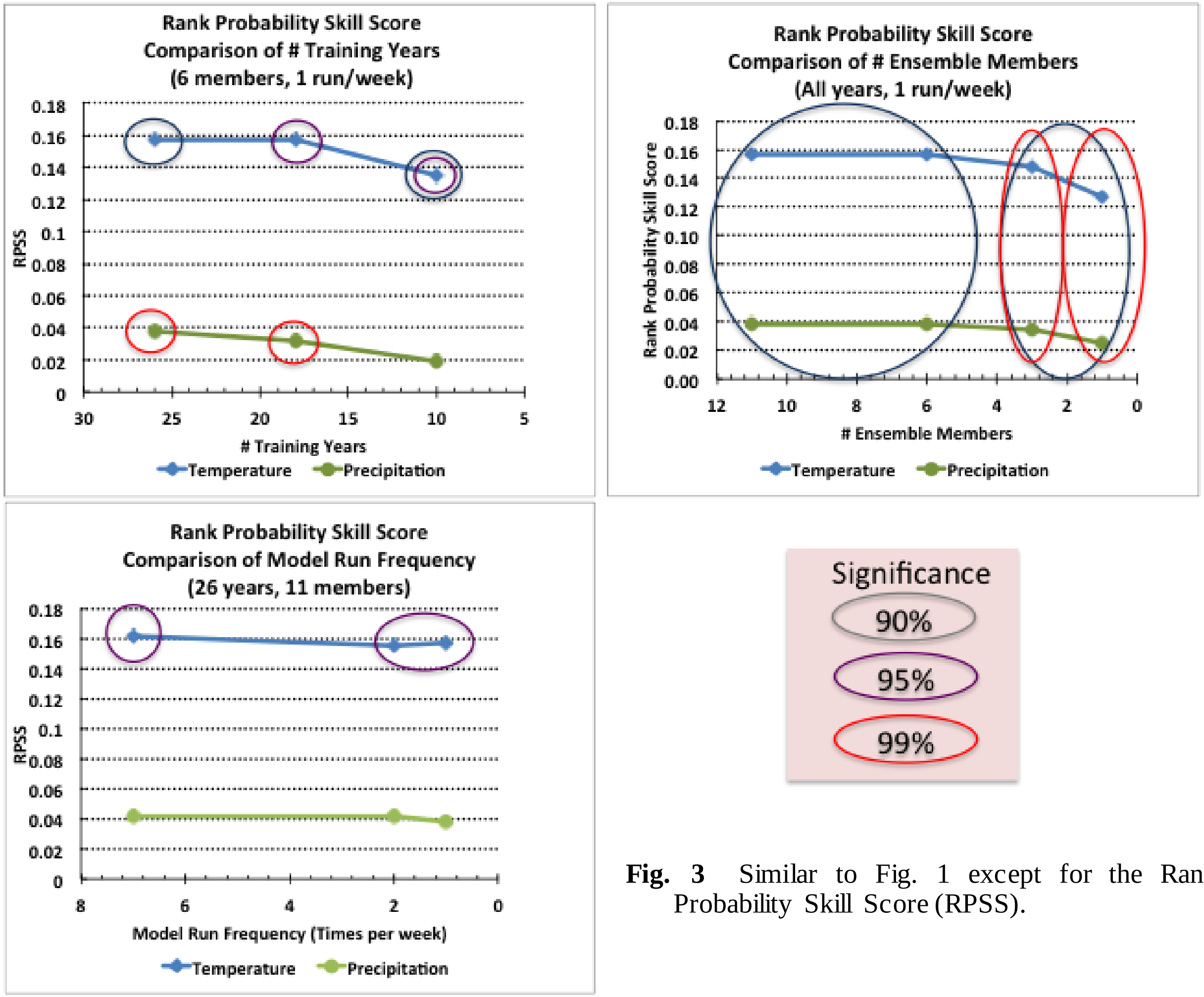

Fig. 3 Similar to Fig. 1 except for the Rank Probability Skill Score (RPSS).

Analysis of line plots of HSS (Fig. 1) shows the most significant drop in skill by changing the number of training years used in the reforecast sample and the least loss of skill from changing the model frequency, for both temperature and precipitation for the shown cases. For temperature, the significant drop in skill (with 6 members and 1 model run/week) occurs when using 10 years instead of 18 (HSS decreases by 1.2), whereas for precipitation the drop occurs when dropping from 26 years to 18 years (HSS decreases by 1.4). These skill differences have a significance level of 95% or greater.

For the selected line plot cases, changing the number of ensemble members only causes a significant drop in skill when using only one member (control run only) for both temperature and precipitation. The model run frequency only causes a noticeable drop in skill when going from 2 runs/week to 1 run/week for precipitation. The HSS of temperature is not greatly impacted by changing the number of ensemble members.

Bar histograms of HSS (Fig. 2) show similar results as the line plots of HSS, with the number of training years impacting the skill most and the number of ensemble members the least. The model run frequency impacts the skill of precipitation forecasts slightly more than temperature. In general, the line plots and

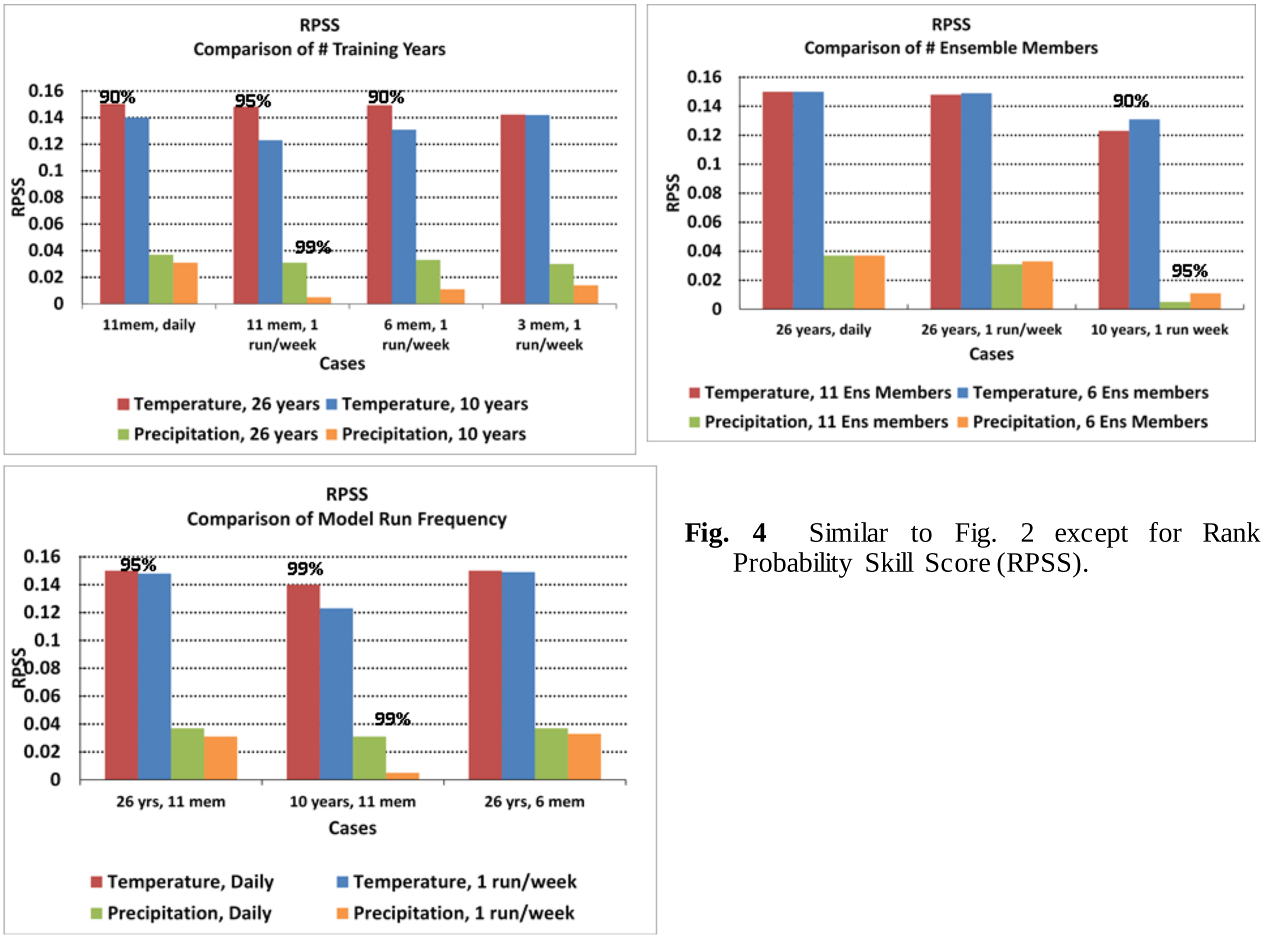

Fig. 4 Similar to Fig. 2 except for Rank Probability Skill Score (RPSS).

histograms of HSS show skill loss of about 2 when using 10 years instead of 26 for both temperature and precipitation.

An interesting feature of the results is that for some cases, a lower configuration of some parameters led to an increase in skill. For temperature, using 6 ensemble members instead of 11 for the case with 26 years and 1 run/week shows a small increase in skill. This is also true for using 1 run/week instead of daily runs (also for temperature). However, these examples do not have a high significance, although some examples shown later on do.

Like the line plots of HSS, the line plots of RPSS also show that the greatest loss in skill results from lower configurations of the number of training years and the least skill loss from changing the model run frequency. Similar to the HSS temperature line plots, the drop in skill occurs when using only 10 years. For precipitation, using 18 years instead of 26 caused a significant decrease in skill. Interestingly, the number of ensemble members shows a noticeable decrease in RPSS when dropping down to 3 members instead of just 1 in the case of the HSS. This would indicate that even though using 3 members does not greatly impact the hit-based aspect of skill assessing the number of correctly forecast, it does affect the ability of the forecast to properly issue the probabilities associated with the forecast.

The bar histograms of RPSS also show the greatest decrease in skill associated with the number of training years. Across the shown cases, the drop in skill from using 10 training years instead of 26 is about 0.02 for both temperature and precipitation. Precipitation has a more varied skill response amongst cases than temperature. Similar to the RPSS line plots, using 6 instead of 11 members across 3 different cases does not yield a significant difference in skill. The impact of model run frequency on skill in the line plot cases was not impressive, although, the cases in the histogram show varying results. The variance of skill difference is especially noticeable for precipitation, where the case of 26 years and 11 members yields a RPSS drop of

0.004 (case shown on line plot) but for 10 years and 11 members, there is a skill drop of 0.025. This exemplifies the importance of evaluating the impact of skill on decreasing training data configurations across various cases, varying the fixed parameters in different ways.

For the case where 10 training years and 1 run/week are used, using 6 ensemble members instead of 11 actually increases the skill by a small amount for temperature and precipitation (0.006 for temperature, 0.004 for precipitation). These results are significant to the 90% level for temperature, and 95% level for precipitation. Improvements in skill by lowering the reforecast sampling configuration would likely be due to overfitting of data, meaning that it is potentially useful to select a lower configuration to improve skill. Doing so, however, may impact the skill in different ways depending on the verification metric or method, such as scoring by separate categories, the type of score, *etc*.

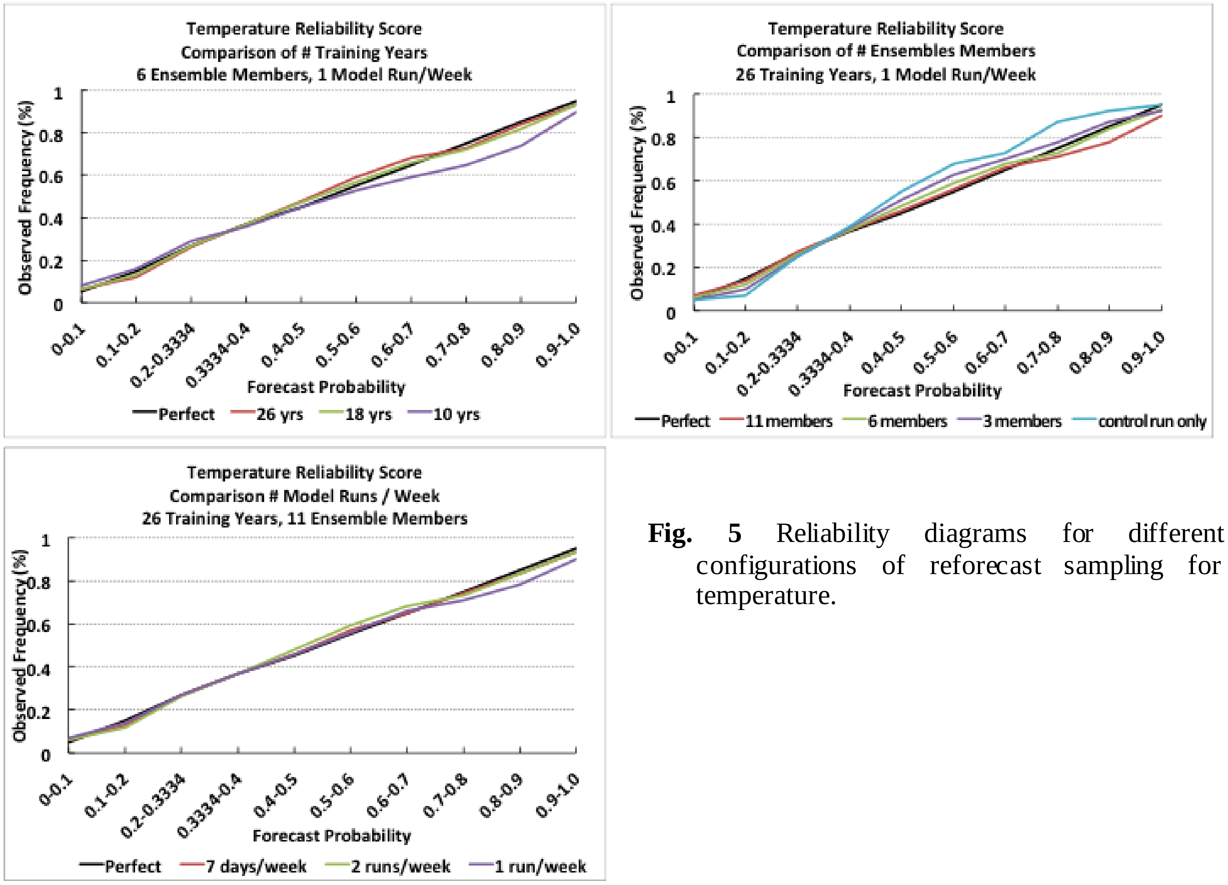

Fig. 5 Reliability diagrams for different configurations of reforecast sampling for temperature.

Reliability diagrams are shown for temperature (Fig. 5) and precipitation (Fig. 6) with the same cases as the line plots. These cases show that temperature forecasts have generally good reliability. The cases that cause the lowest reliability is using one ensemble member only (the control run) and using 10 training years. Using only the control run leads to forecast probabilities that are too low (for forecast probabilities 40% or greater) and using 10 years leads to forecast probabilities that are too high (for forecast probabilities of 60% or greater). The forecast probabilities that are too low or high are reflected by curves on the reliability diagram that lie above and below the perfect skill line, respectively.

Overall, the reliability of the precipitation forecasts (for the selected cases) show worse reliability than temperature, which is to be expected due to the inherent nature of it being a harder quantity to forecast. There is greater spread amongst the precipitation reliability curves across the cases compared to temperature, indicating that precipitation is more sensitive to reforecast sampling than temperature. These diagrams indicate the greatest decrease in skill results from using only 10 members and only the control run, while the model run frequency impacts skill the least, which is similar to the behavior of the reliability scores of temperature.

It is evident that the reliability curves start deviating from the perfect score line at lower probabilities in precipitation than temperature which may indicate that there is a greater range of forecast probabilities that are less reliable for precipitation. Regardless of these differences, the overall behavior of the skill seems to react in a similar manner for both temperature and precipitation when changing the parameters of the reforecast sampling.

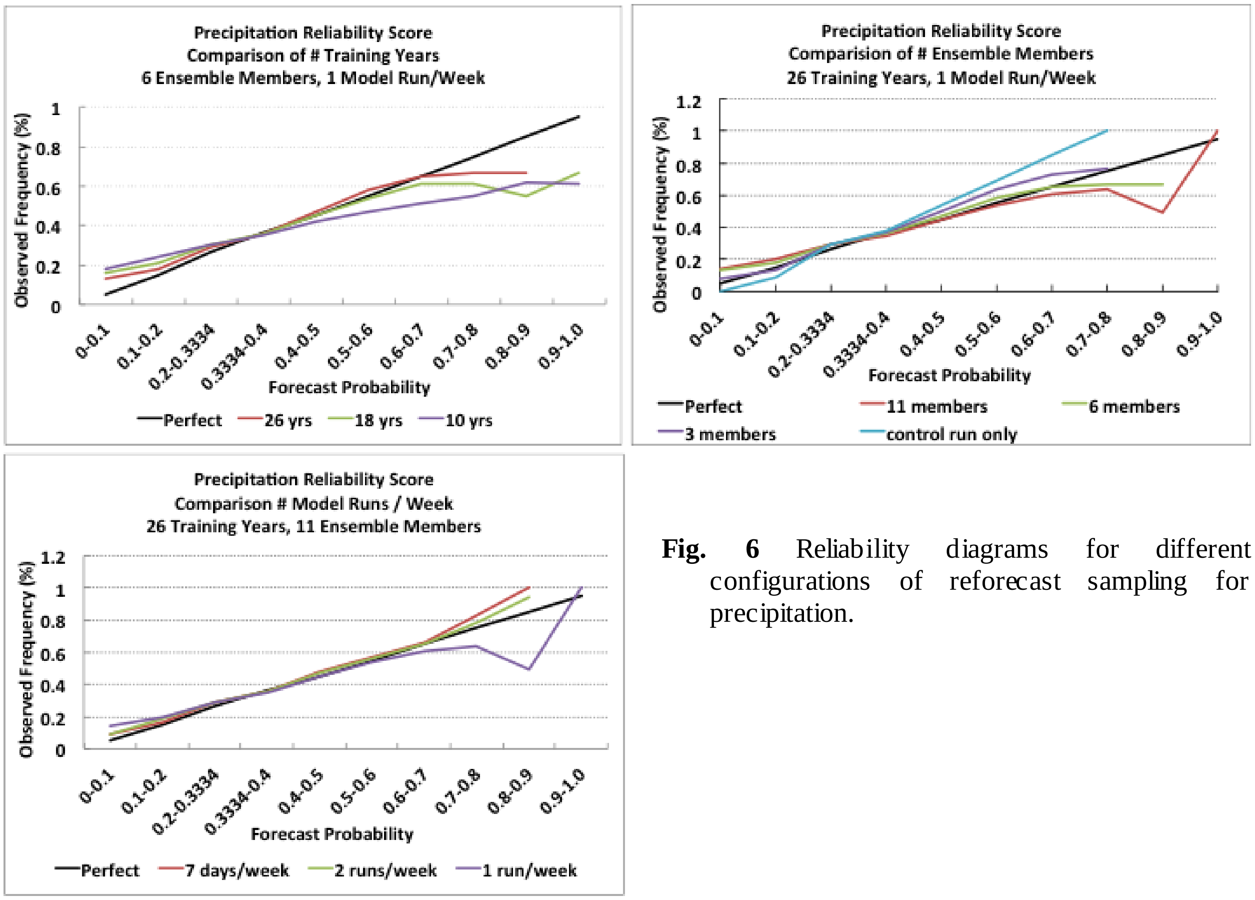

Fig. 6 Reliability diagrams for different configurations of reforecast sampling for precipitation.

4. Conclusions

Evaluation of skill scores of the week-2 reforecast tool indicate that temperature and precipitation forecasts are most sensitive to the number of training years of reforecasts used in the calibration and the least sensitive to the model run frequency. In general, precipitation forecasts are more sensitive to decreasing configurations of reforecasts than temperature.

It is important to assess the skill of forecasts using different score types. Skill metrics that assess forecast quality based on probabilities were impacted differently than those that assess the number of correctly guessed forecasts. Using 3 members or less of reforecasts did not impact the Heidke skill score, but it did noticeably decrease the RPSS.

The skill of precipitation forecasts is more sensitive to decreasing configurations of reforecasts than temperature, especially when evaluating the probabilistic aspect of skill. This shows that the forecast skill of different atmospheric variables may be sensitive to the configuration of different reforecast parameters. Therefore, determining the minimum required reforecasts for reforecast production should be driven by the atmospheric variable (of the variables desired to be forecast) that exhibits the most skill sensitivity to changing configurations of reforecast sampling. Despite the differences between temperature and precipitation regarding skill sensitivity, the type of impact to each are similar (*e.g.* using the control run only leads to forecast probabilities that are too low and using only 10 training years produces probabilities that are too high).

Even though decreasing the configuration of reforecasts may lead to some forecasts' decrease in skill, it is also evident that selective lower configurations can still produce skillful week-2 forecasts with minimal skill loss. Dropping down from 11 members to 6 and daily reforecasts to 1 model run per week only causes a minor decrease in skill, if sufficient training years are used.

CPC's recommendation for a lower configuration of reforecast sampling without significant week-2 forecast skill loss is to produce as many years as possible with 6 ensemble members and 1 run/week (weekly). In this case, the maximum years available for this study was 26 years, but 30 years would be desirable since it would be consistent with the standard CPC follows for their climatology and may allow forecasts to see further improvements in skill than the reforecast tool currently has. This configuration seems optimal from these evaluations based on significantly reducing the required resources needed to produce reforecasts without causing a great drop in skill for week-2 temperature and precipitation forecasts. Compared to real-time ensemble forecasts with 21 members per cycle and 4 cycles per day, weekly 6 member (5 members plus control run) reforecasts for 30 years would cost approximately 26% of the computing of the real-time ensemble reforecasts.

References

Unger, David A., Huug van den Dool, Edward O'Lenic, Dan Collins, 2009: Ensemble Regression. *Mon. Wea. Rev.*, **137**, 2365–2379.