Analog Probabilistic Precipitation Forecasts Using GEFS Reforecasts and Climatology-Calibrated Precipitation Analyses

Thomas M. Hamill,1 Michael Scheuerer,2 and Gary T. Bates2

1 NOAA Earth System Research Lab, Physical Sciences Division, Boulder, Colorado

2 CIRES, University of Colorado, Boulder, Colorado

Submitted to Monthly Weather Review

as an expedited contribution

revised

20 March 2015

Corresponding author:
Dr. Thomas M. Hamill
NOAA Earth System Research Lab
Physical Sciences Division
R/PSD 1, 325 Broadway
Boulder, CO 80305
Tom.Hamill@noaa.gov
Phone: (303) 497-3060
Telefax: (303) 497-6449
ABSTRACT

Analog post-processing methods have previously been applied using precipitation reforecasts and analyses to improve probabilistic forecast skill and reliability. A modification to a previously documented analog procedure is described here that produces highly skillful, statistically reliable precipitation forecast guidance at a 1/8th-degree grid spacing. These experimental probabilistic forecast products are available via the web in near real-time.

The main changes to the previously documented analog algorithm were as follows: (a) use of a shorter duration (2002-2013) but smaller grid spacing, higher-quality time series of precipitation analyses for training and forecast verification, the Climatology Calibrated Precipitation Analysis; (b) increased training sample size using data from 19 supplemental locations, chosen for their similar precipitation analysis climatologies and terrain characteristics; (c) selection of analog dates for a particular grid point based on the similarity of forecast characteristics at that grid point rather than similarity in a neighborhood around that grid point; (d) using an analog rather than a rank-analog approach; (e) varying the number of analogs used to estimate probabilities from a smaller number (50) for shorter-lead forecasts to a larger number (200) for longer-lead events; (f) spatial Savitzky-Golay smoothing of the probability fields. Special procedures were also applied near coasts and country boundaries to deal with data unavailability outside of the US while smoothing.

The resulting forecasts are much more skillful and reliable than raw ensemble guidance across a range of event thresholds. The forecasts are not nearly
as sharp, however. The use of the supplemental locations is shown to especially improve the skill of short-term forecasts during the winter.
1. **Introduction.**

Previous studies have shown that probabilistic forecasts of precipitation can be significantly improved by post-processing with reforecasts (e.g., Hamill et al. 2006, hereafter H06; Hamill et al. 2013, hereafter H13; Hamill and Whitaker 2006, hereafter HW06). The real-time forecast is adjusted using a long time series of past forecasts and associated precipitation analyses. Appealing for its simplicity was the “analog” procedure used in these studies. For a given location, dates in the past were identified that had reforecasts similar to today’s forecast. An ensemble was formed from the observed or analyzed precipitation amounts on the dates of the chosen analogs, and probabilities were estimated from the ensemble relative frequency. Maps of precipitation probabilities were constructed by repeating the procedure across the model grid points.

A challenge with analog procedures used in these previous studies was their inability to find many close-matching forecasts when today’s precipitation forecast amount was especially large, even with a long training data set. The method as previously documented used the data surrounding grid point of interest but did not supplement the training data set with observation and forecast data centered on other locations. The benefit of this location-specific approach was that if the model’s systematic errors varied greatly with location, it corrected for these, as shown in H06. One disadvantage was that if there were not many prior forecasts with similarly extreme precipitation, then the selected analogs were biased toward precipitation forecasts with less extreme forecast values and typically lighter
analyzed precipitation. Consequently, the forecast procedure did not often produce high probabilities of extreme events.

Another possible disadvantage of the forecast products demonstrated in these previous studies was that the associated precipitation analyses were in each case from the North American Regional Reanalysis (NARR, Mesinger et al. 2006). Several studies have identified deficiencies with this data set (e.g., West et al. 2007, Bukovsky and Karoly 2009). We have also noted a significant dry bias in the NARR over the northern Great Plains during the winter season. There is now an alternative data set that covers the contiguous US (CONUS) and that utilizes both gauge and adjusted radar-reflectivity data, the Climatology-Calibrated Precipitation Analysis (CCPA; Hou et al. 2014). Data is available from 2002-current. While this time period is shorter than the 1985-current time span of the most recent reforecast (H13), the availability of higher-resolution, more accurate precipitation analysis data has led us to consider whether useful products could be generated with this data set.

This article briefly describes modifications to previously documented analog forecast procedures. What adjustments will allow it to provide improved probabilistic forecasts while using a shorter time series of analyses? We describe a series of changes to the analog algorithm and show that the resulting analog probabilistic forecasts are skillful, somewhat more sharp, and reliable. Since the statistically post-processed guidance provide a significant improvement over probabilities from the raw Global Ensemble Forecast System (GEFS) forecast data, we are also making experimental web-based guidance available in near real time.
during the next few years; this guidance can be obtained from


2. Methods and data.

a. Reforecast data, observational data, and verification methods.

In this study we considered 12-hourly accumulated precipitation forecasts
during the 2002 to 2013 period for lead times up to +8 days. Precipitation analyses
were obtained on a ~1/8-degree grid from the CCPA data set of Hou et al. (2014).
Probabilistic forecasts were produced at this ~1/8-degree resolution over the
CONUS. All of the forecast data used in this project were obtained from the second-
generation GEFS reforecast data set, described in H13. Ensemble-mean
precipitation and total-column ensemble-mean precipitable water were used in the
analog procedure. GEFS data was extracted (for precipitation) on the GEFS's native
Gaussian grid at ~1/2-degree resolution in an area surrounding the CONUS.
Precipitable-water forecasts, which were archived on a 1-degree grid, were
interpolated to the native Gaussian grid before input to the analog procedure.
Forecasts were cross validated; for example, 2002 forecasts were trained
using 2003-2013 data. For the production of forecasts in a given month, the training
data used that month and the surrounding two months, e.g., January forecasts were
trained with December-January-February data. Was the use of future data in the
cross-validation procedure a source of unrealistic skill of these forecasts? As shown
in Baxter et al. (2014, Fig. 5 therein), the inter-annual variability of skill in the
southeast US was larger than the systematic changes from 2002 to 2013. This
suggests that the use of future forecasts in the cross-validation procedure probably
did not result in a large over-estimation of forecast skill for the earlier years.
One of the controls against which the new method was compared were the
raw event probabilities generated from the 11-member GEFS reforecast ensemble,
bi-linearly interpolated to the 1/8-degree grid.
Verification methods included reliability diagrams and Brier Skill Scores
computed in the conventional way (Wilks 2006, eqs. 7.34 and 7.35, Hamill and Juras
2006), with climatology providing the reference probabilistic forecasts. Maps of
Brier Skill Scores were also generated for the CONUS. These were produced by
accumulating the probabilistic forecasts’ and climatological forecasts’ average of
squared error at that grid point across all years and all months prior to the
calculation of skill. Because of the extremely large sample size, confidence intervals
for the skill differences (very small; see HW06) were not included on the plots.

b. Rank analog forecast procedure as a control.
A revised “rank-analog” approach served as another standard of comparison
for the newer, somewhat more involved analog methodology described in section
2.c below. For the most part, the rank-analog approach was a hybrid of the
techniques that have previously been shown to work well, described in sections
3.b.6 and 3.b.8 of HW06. This control rank-analog methodology was further
updated in the following respects:

- As with the rank-analog algorithm of HW06, the rank of the forecast for a
sorted forecasts at the same set of grid points for each date in the training data set.

In evaluating which forecasts were closest to today’s forecast, the difference between forecasts was calculated as 70% of the absolute difference of the precipitation forecast ranks and 30% of the absolute difference in precipitable water forecast ranks averaged over the set of grid points, following HW06. As shown therein, with the exception of warm-season probability of precipitation, there was minimal sensitivity to the chosen weight between precipitation and precipitable water. A more precise definition of the forecast difference is as follows:

let $S$ be the set of grid points in a region surrounding the current grid point of interest. Let $tc$ be the current date, and let $t$ be another date from the set of dates $T$ whose forecast data will be compared against the forecast at $tc$. As indicated previously, by cross validation $tc \notin T$. Define $r_{pr_s}^{tc}$ as the rank of the current forecast precipitation amount at time $tc$ and at grid point $s$ from a combined set with the training data at $s$. Similarly, $r_{pw_s}^{tc}$ is the associated rank of the current total-column precipitable water forecast. Then the difference in ranks for date $t$ was calculated as

$$d_t = \sum_{s=1}^{S} [|0.7 \times (r_{pr_s}^{tc} - r_{pr_s}^{t})| + |0.3 \times (r_{pw_s}^{tc} - r_{pw_s}^{t})|].$$  \hspace{1cm} (1)$$

The chosen date $t$ was then simply the date in $T$ that had the minimum difference. Once this date was selected, it was omitted from further consideration.

- The size of the search region for pattern matching of forecasts was allowed to vary with forecast lead time, inspired by the results of testing the method described in 3.b.9 of HW06. Specifically, let $t_e$ denote the end of the forecast precipitation accumulation period in hours, and let $\delta$ denote the box width in units
of numbers of grid points on the ~ 1/2-degree Gaussian grid. If $t_{e} \leq 48$, then $\delta = 5$; if $48 < t_{e} \leq 96$, then $\delta = 7$; if $96 < t_{e} \leq 132$, then $\delta = 9$; if $132 < t_{e}$, then $\delta = 11$.

- The number of analogs used in the generation of probabilities was allowed to vary as a function of the forecast lead time and how unusual was the precipitation forecast in question, measured in terms of its percentile relative to the climatological distribution of forecasts ($q_{f}$). Let $n_{a}$ be the number of analogs used. If the end period for the forecast precipitation was $> 48$ h, then when $q_{f} < 0.75$, $n_{a} = 100$; when $0.75 \leq q_{f} < 0.9$, $n_{a} = 75$; when $0.9 \leq q_{f} < 0.95$, $n_{a} = 50$; when $q_{f} \geq 0.95$, $n_{a} = 25$. If the end period for the forecast $\leq 48$ h, then when $q_{f} < 0.75$, $n_{a} = 50$; when $0.75 \leq q_{f} < 0.9$, $n_{a} = 40$; when $0.9 \leq q_{f} < 0.95$, $n_{a} = 30$; when $q_{f} \geq 0.95$, $n_{a} = 20$. This dependence of analog size on forecast lead time and unusualness of the forecast with respect to the climatology was inspired by the results of Fig. 7 and associated discussion in H06. This showed that fewer analogs provided the best skill for shorter lead times and for heavy-precipitation events; more analogs were desirable at longer leads and for more common light- or no-precipitation events. The values do not correspond exactly with the optimal values from H06 in part because the length of the training data set was somewhat shorter here (11 years with cross validation).

c. New analog procedure with additional training data from supplemental locations.

We now describe an update to the basic analog (hereafter, simply “analog”) procedure described in section 3.a.3 of HW06. This revised procedure was evaluated here against the rank-analog procedure described in section 2.b, and was
used in the generation of our real-time web graphics. The following modifications were made:

- Analogs were chosen not by finding a forecast pattern match in an area surrounding the analysis grid point of interest, but rather by using only the forecast data specifically at a grid point, as in Delle Monache et al. (2013). With this modification, data from other supplemental grid points, described below, could be used as additional training samples. In large part, the reason for not using a rank analog with a pattern match over an area was computational efficiency; with many extra supplemental locations under consideration, matching forecasts at points was much faster than matching forecasts over regions encompassing many grid points.

- The number of analogs used in the computation of the probabilities varied with forecast lead time. The number of analogs was defined as follows: if the end period $t_e$ for the forecast precipitation was $\leq 24$ h, then $n_a=50$; if $24 < t_e \leq 48$ h, $n_a=75$; if $48 \leq t_e < 96$ h, $n_a=100$; if $96 \leq t_e < 120$ h, $n_a=150$; if $t_e \geq 120$ h, $n_a=200$.

However, unlike the rank-analog method described above, the number of analogs was not allowed to vary based on the unusualness of today’s forecast; it was judged that ample training data was available in most situations, given the extra data from the 19 supplemental locations.

- In the selection of analog dates, the interpolated forecast for a particular date of interest and analysis grid point $(i,j)$ was compared against interpolated forecasts at $(i,j)$ for each date in the training data set. In evaluating which forecasts were closest to today’s forecast, the difference between forecasts was calculated as $70\%$ of the absolute difference of the precipitation forecasts and $30\%$ of the
absolute difference in precipitable water forecasts. That is, let $pr_{i,j}^{fc}$ be the forecast precipitation amount at the grid point $(i,j)$ and the current date, and $pw_{i,j}^{fc}$ be total-column precipitable water. Then the difference $d_t$ at a different date $t$ was

$$d_t = |0.7 \times (pr_{i,j}^{fc} - pr_{i,j}^t) + 0.3 \times (pw_{i,j}^{fc} - pw_{i,j}^t)|.$$ (2)

Note that here the ranks of the precipitation values were not compared, as in the prior algorithm, but rather the raw forecasts values.

- The interpolated forecast for a particular date of interest and grid point $(i,j)$ was also compared against interpolated forecasts at 19 other supplemental locations $(i_s,j_s)$ on other dates. When the closest match was found to occur with data at one of these supplemental locations, then the analysis from this supplemental location on this date was used as an analog member. That supplemental member and date were then excluded from further consideration.

- The 19 supplemental locations were determined for each grid point based upon the similarity of the observed climatology and the similarity of terrain characteristics. There were also constraints on a minimum distance between supplemental locations and a penalty for distance between points. The specific methodology of defining supplemental locations is described in the online appendix A. An example of the selected supplemental locations and their relation to the local climatology is shown in Fig. 1.

- Once probability forecasts were generated from the ensemble of analyzed states on the dates of the selected forecast analogs, the probability forecasts were smoothed using a 2-D Savitzky-Golay smoother with a window size of 9 grid points.
and using a third-order polynomial. The details of this smoother are also described in the online appendix A.

Which of the changes above were significant and which were more minor?

Not considering supplemental locations, the use of the analog with point data vs. the rank analog with surrounding-area data decreased skill somewhat (not shown).

However, the inclusion of supplemental training data had an even bigger positive impact and provided overall the largest impact on skill and reliability. The variable number of analog members with forecast lead produced a smaller improvement relative to using the same number at all leads. The smoothing did not affect the reliability or skill much, but the resulting forecasts were much more visually appealing. Online Appendix A provides an example of the before vs. after smoothing difference.

3. Results.

Figures 2 and 3 show Brier Skill Scores as a function of forecast lead time for the > 1 mm (12 h)$^{-1}$ event and the > 25 mm (12 h)$^{-1}$, respectively. Skill scores for other event thresholds are presented in online appendix B. While both rank analog and analog forecasts provided a significant improvement with respect to the raw guidance, the skills of the warm-season forecasts at shorter leads from the newer analog method for the > 1 mm event were slightly lower skill than those of the rank-analog method. This was likely because the > 1 mm event was not an especially rare event at most locations, so the increased sample size with the new analog method did not compensate for the other relative advantages of using a rank-analog rather than a straight analog approach. Considering the skill for the > 25 mm event
in Fig. 3, the new analog procedure did provide a skill improvement, especially for shorter-lead forecasts during the cool season. In these circumstances, the day +2 analog forecasts with supplemental locations were more skillful than the day +1 rank analog forecasts, and both were notably higher in skill than the raw ensemble. Why was there greater improvement of heavy precipitation forecasts with the new analog procedure in winter? Though not confirmed, we hypothesize that in winter there was higher intrinsic skill of the forecasts than in summer, due to the different phenomena driving precipitation with their different space and time scales: synoptic-scale ascent in mid-latitude winter cyclones, thunderstorms during the summer season. Further, in wintertime, there were larger fluctuations of the probabilities about their long-term climatological mean with meaningful signal. Thus the additional samples helped refine the estimates of $O|F$, the conditional distribution of observations given the forecast (HW06, eq. 3), thereby improving the probabilistic forecast, despite the lack of pattern matching used in the rank-analog approach.

Figure 4 shows maps of Brier skill scores for the $>1$ mm event at the 60-72-h lead time. There was little difference between the two analog forecasts, consistent with Fig. 2. Both were more skillful than the raw ensemble, which had BSS < 0 over a significant percentage of the country, in part due to sampling error (Richardson 2001) but mostly due to systematic errors and sub-optimal treatment of model uncertainty in the GEFS. Skill for all methods was largest in mountainous areas along the US West Coast, with the predictable phenomena of the flow from mid-latitude cyclones impinging upon the stationary topography. Figure 5 shows maps
of skill for the > 25 mm event at the 60-72-h lead time. There appeared to be a
general improvement in skill across the country for the analog with supplemental
locations. Again, raw ensembles were notably unskillful across drier regions of the
US but competitive in a few select locations in the Sierra Nevada mountain range.
Maps for other forecast lead times and thresholds are provided in online Appendix B.
The resulting post-processed forecast guidance was consistently reliable, too.
Figure 6 provides reliability diagrams for the three methods for > 25 mm and 60-72
h forecast leads; again, see appendix B for more diagrams at other leads and event
thresholds. Both analog methods were quite reliable, though the analog with
supplemental locations had somewhat more forecasts issuing high-probabilities
(greater sharpness). Both analog methods were much less sharp than the raw
forecast guidance but more reliable. Why was the analog method with
supplemental locations sharper? This was because the extra training sample size
permitted the identification of closer analogs than with the rank-analog approach.
As noted in HW06, a general challenge with the analog or rank-analog forecasts
(therein without supplemental location data) of extreme events was their inability
to find many forecasts dates with amounts that were similar in magnitude.

4. Discussion and conclusions

This article has demonstrated an improved method for post-processing that
provides dramatically improved guidance of probabilistic precipitation when paired
with a reforecast data set of sufficient length and precipitation analyses of sufficient
quality. This article provides additional evidence to support the assertion that the
regular production of weather reforecasts will help with the objective definition of high-impact event probabilities.

Though the use of supplemental locations was shown to provide significant improvement to heavy precipitation forecast calibration, our examination of possible methods for choosing the location and number of supplemental location data was far from systematic. The methods for the selection of these locations deserves further study.

This method may provide a useful benchmark for comparison of other methods. Whereas the analog method here has been shown to work well with larger reforecast data sets, these are not always available. We anticipate subsequent studies will compare the efficacy of analog methods with respect to other (e.g., parametric) post-processing methods, including when using much smaller training sample sizes. In this way we hope to understand whether the choice of a preferred post-processing algorithm is robust from small to large training sample sizes.

Acknowledgments:

This research was supported by a NOAA US Weather Program grant as well as funding from the National Weather Service Sandy Supplemental project. The reforecast data set was computed at the US Department of Energy's (DOE) National Energy Research Computing Center, a DOE Office of Science user facility.
References


analysis. *J. Hydrometeor*, **15**, 2542-2557. doi:

http://dx.doi.org/10.1175/JHM-D-11-0140.1


**Figure captions**

**Figure 1.** Illustration of the location of supplemental locations and their dependence on the analyzed precipitation climatology. Colors denote the 95th percentile of the analysis distribution for the month of January, based on 2002-2013 CCPA data. Supplemental data locations are also shown. The larger symbols indicate sample locations where supplemental data is sought, and the smaller symbols indicate the chosen supplemental locations.

**Figure 2:** Brier skill scores for the > 1 mm (12 h)\(^{-1}\) event over a range of lead times as a function of the month of the year. (a) Skills of forecasts from the new analog method with 19 supplemental locations; (b) skills of forecasts from the older rank-analog method for comparison; (c) skills of forecasts from the 11-member raw ensemble guidance.

**Figure 3:** As in Fig. 2, but for the > 25 mm (12 h)\(^{-1}\) event. The climatology is computed separately for each month and each ~1/8-degree grid point location.

**Figure 4:** Maps of yearly 60-72 h forecast Brier Skill Scores, for probabilistic forecasts of the > 1 mm (12 h)\(^{-1}\) event, generated from (a) analog forecasts with 19 supplemental locations, (b) rank analog forecast with no supplemental locations, and (c) 11-member raw ensemble.

**Figure 5:** As in Fig. 4, but for > 25 mm event.

**Figure 6:** Reliability diagrams for the > 25 mm event for 60- to 72-h forecasts. (a) analog forecasts with 19 supplemental locations, (b) rank analog forecast with no supplemental locations, and (c) 11-member raw ensemble.
Figure 1. Illustration of the location of supplemental locations and their dependence on the analyzed precipitation climatology. Colors denote the 95th percentile of the analysis distribution for the month of January, based on 2002-2013 CCPA data. Supplemental data locations are also shown. The larger symbols indicate sample locations where supplemental data is sought, and the smaller symbols indicate the chosen supplemental locations.
Figure 2: Brier skill scores for the > 1 mm (12 h)^{-1} event over a range of lead times as a function of the month of the year. (a) Skills of forecasts from the new analog method with 19 supplemental locations; (b) skills of forecasts from the older rank-analog method for comparison; (c) skills of forecasts from the 11-member raw ensemble guidance. (d) skill difference, analog minus rank analog; (e) skill difference, rank analog minus raw.
**Figure 3:** As in Fig. 2, but for the event of greater than > 25 mm (12 h)^{-1}.
Figure 4: Maps of yearly 60-72 h forecast Brier Skill Scores, for probabilistic forecasts of the > 1 mm 12 h⁻¹ event, generated from (a) analog forecasts with 19 supplemental locations, (b) rank analog forecast with no supplemental locations, and (c) 11-member raw ensemble.
Figure 5: As in Fig. 4, but for > 25 mm (12 h)^{-1} event.
Figure 6: Reliability diagrams for the > 25 mm (12 h)^{-1} event for 60- to 72-h forecasts. (a) analog forecasts with 19 supplemental locations, (b) rank analog forecast with no supplemental locations, and (c) 11-member raw ensemble.