An Empirical Cumulative Density Function Approach to Defining Summary NWP Forecast Assessment Metrics

ROSS N. HOFFMAN

NOAA/Atlantic Oceanographic and Meteorological Laboratory, and Cooperative Institute for Marine and Atmospheric Studies, University of Miami, Miami, Florida

SID-AHMED BOUKABARA

NOAA/NESDIS/Center for Satellite Applications and Research, College Park, Maryland

V. KRISHNA KUMAR AND KEVIN GARRETT

Riverside Technology Inc., Joint Center for Satellite Data Assimilation, and NOAA/NESDIS/Center for Satellite Applications and Research, College Park, Maryland

SEAN P. F. CASEY

NOAA/Atlantic Oceanographic and Meteorological Laboratory, and Cooperative Institute for Marine and Atmospheric Studies, University of Miami, Miami, Florida

ROBERT ATLAS

NOAA/Atlantic Oceanographic and Meteorological Laboratory, Miami, Florida

(Manuscript received 15 July 2016, in final form 18 October 2016)

ABSTRACT

The empirical cumulative density function (ECDF) approach can be used to combine multiple, diverse assessment metrics into summary assessment metrics (SAMs) to analyze the results of impact experiments and preoperational implementation testing with numerical weather prediction (NWP) models. The main advantages of the ECDF approach are that it is amenable to statistical significance testing and produces results that are easy to interpret because the SAMs for various subsets tend to vary smoothly and in a consistent manner. In addition, the ECDF approach can be applied in various contexts thanks to the flexibility allowed in the definition of the reference sample.

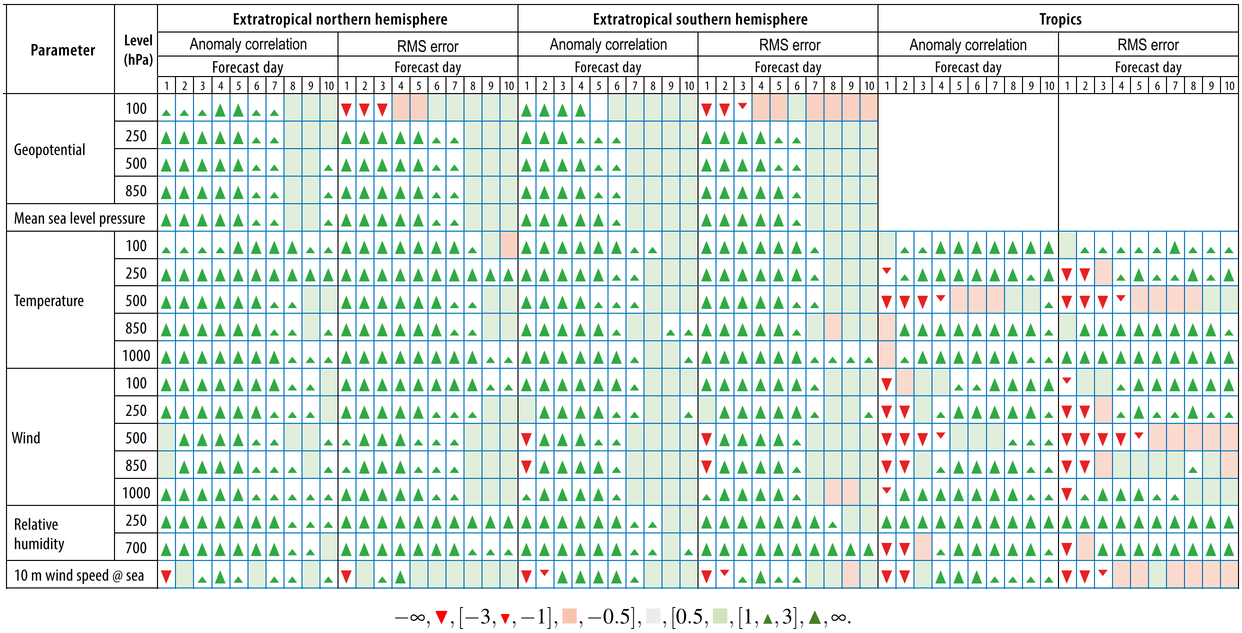
The interpretations of the examples presented here of the impact of potential future data gaps are consistent with previously reported conclusions. An interesting finding is that the impact of observations decreases with increasing forecast time. This is interpreted as being caused by the masking effect of NWP model errors increasing to become the dominant source of forecast error.

1. Introduction

A welter of quantitative assessment metrics are produced by modern numerical weather prediction (NWP) data assimilation and forecast systems. Boukabara et al. (2016, hereafter BGK) introduced the overall forecast score (OFS) as a mathematically rigorous, yet simple, approach to compositing large collections of diverse assessment metrics using normalized scores combining different variables, levels, forecast times, and metrics. We present in this study an alternative approach to computing a composite score that relies on an empirical cumulative distribution function (ECDF) to normalize the assessment metrics. In simple terms, the ECDF normalization of a particular forecast metric is the fraction of forecasts that are worse than the given forecast. A principal advantage of the ECDF approach is in assigning confidence intervals to the composite score. In this discussion, the assessment metrics are the forecast anomaly correlation (AC) and the forecast

Corresponding author e-mail: Ross N. Hoffman, ross.n.hoffman@noaa.gov

DOI: 10.1175/MWR-D-16-0271.1



Parameter	Level (hPa)	Extratropical northern hemisphere																				Extratropical southern hemisphere																				Tropics																			
		Anomaly correlation										RMS error										Anomaly correlation										RMS error										Anomaly correlation										RMS error									
		Forecast day										Forecast day										Forecast day										Forecast day										Forecast day										Forecast day									
		1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10
Geopotential	100	▴	▴	▴	▲	▲	▴	▴	🟩	🟩	🟩	▼	▼	▼	🟥	🟥	🟩	🟩	🟩	🟩	🟩	▲	▲	▲	▲		🟩	🟩	🟩	🟩	🟩	▼	▼	▾	🟥	🟥	🟩	🟥	🟥	🟥	🟥																				
	250	▲	▲	▲	▲	▲	▴	▴	🟩	🟩	🟩	▲	▲	▲	▲	▲	▴	▴	🟩	🟩	🟩	▲	▲	▲	▴	▴	▴	🟩	🟩	🟩	🟩	▲	▲	▲	▲	▴	▴	🟩	🟩	🟩	🟩																				
	500	▲	▲	▲	▲	▲	▴	▴	🟩	🟩	▴	▲	▲	▲	▲	▲	▴	▴	🟩	🟩	🟩	▲	▲	▲	▲	▴	▴	🟩	🟩	🟩	🟩	▲	▲	▲	▲	▴	▴	🟩	🟩	🟩	🟩																				
	850	▲	▲	▲	▲	▲	▴	▴	🟩	🟩	▴	▲	▲	▲	▲	▲	▴	▴	🟩	🟩	🟩	▲	▲	▲	▲	▴	▴	🟩	🟩	🟩	🟩	▲	▲	▲	▲	▲	▴	🟩	🟩	🟩	🟩																				
Mean sea level pressure		▲	▲	▲	▲	▲	▴	▴	🟩	🟩	▴	▲	▲	▲	▲	▲	▴	▴	🟩	🟩	🟩	▲	▲	▲	▲	▲	▴	🟩	🟩	🟩	🟩	▲	▲	▲	▲	▲	▴	🟩	🟩	🟩	🟩																				
Temperature	100	▴	▴	▴	▴	▲	▲	▲	▲	▴	▴	▲	▲	▲	▲	▲	▲	▲	▴	🟩	🟥	▲	▲	▲	▲	▲	▲	▴	▴	🟩	🟩	▲	▲	▲	▲	▲	▲	▴	🟩	🟩	🟩	🟩	▴	▴	▲	▲	▲	▲	▲	▲	▲	🟩	▴	▴	▴	▴	▴	▲	▴	▴	▴
	250	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▴	▴	🟩	🟩	🟩	▲	▲	▲	▲	▲	▲	▴	🟩	🟩	🟩	▾	▴	▲	▲	▲	▲	▲	▲	▴	▲	▼	▼	🟥	▴	▲	▴	▴	▲	▴	▲
	500	▲	▲	▲	▲	▲	▲	▴	▴	🟩	🟩	▲	▲	▲	▲	▲	▲	▴	▴	🟩	🟩	▲	▲	▲	▲	▴	▴	▴	🟩	🟩	🟩	▲	▲	▲	▲	▲	▴	▴	🟩	🟩	🟩	▼	▼	▼	▾	🟥	🟥	🟥	🟩	🟩	▴	▼	▼	▼	▾	🟥	🟥	🟥	🟥	🟩	🟩
	850	▲	▲	▲	▲	▲	▲	▴	▴	🟩	▴	▲	▲	▲	▲	▲	▲	▴	▴	🟩	🟩	▲	▲	▲	▲	▲	▴	🟩	🟩	▴	▴	▲	▲	▲	▲	▲	▴	🟩	🟥	🟩	🟩	🟥	▲	▲	▲	▲	▲	▲	▲	▲	▴	🟩	▲	▲	▲	▲	▲	▲	▲	▲	▴
	1000	▲	▲	▲	▲	▲	▲	▲	▴	▴	▴	▲	▲	▲	▲	▲	▲	▲	▲	▴	▴	▲	▲	▲	▲	▲	▲	▴	🟩	🟩	▴	▲	▲	▲	▲	▲	▲	▴	▴	▴	▴	🟥	▴	▲	▲	▲	▲	▲	▲	▴	▴	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Wind	100	▲	▲	▲	▲	▲	▲	▲	▴	▴	🟩	▲	▲	▲	▲	▲	▲	▲	▲	▴	▴	▲	▲	▲	▲	▲	▲	▴	🟩	🟩	🟩	▲	▲	▲	▲	▲	▲	▴	🟩	🟩	🟩	▼	🟥	🟩	🟩	▴	▴	▲	▲	▲	▲	▾	🟩	🟩	▴	▲	▲	▲	▲	▲	▲
	250	▲	▲	▲	▲	▲	▲	▴	▴	▴	🟩	▲	▲	▲	▲	▲	▴	▴	▴	🟩	🟩	🟩	▲	▲	▲	▲	▴	▴	🟩	🟩	▴	🟩	▲	▲	▲	▲	▲	▴	🟩	🟩	▴	▼	▼	🟩	▴	▲	▲	▲	▲	▲	▴	▼	▼	🟥	▴	▲	▴	▴	▲	▴	▴
	500	🟩	▲	▲	▲	▲	▴	▴	🟩	🟩	▴	▴	▲	▲	▲	▴	▴	▴	🟩	🟩	🟩	▼	▲	▲	▲	▴	▴	🟩	🟩	🟩	🟩	▼	▲	▲	▲	▲	▴	🟩	🟩	🟩	🟩	▼	▼	▼	▾	🟩	🟩	🟩	▴	▴	▴	▼	▼	▼	▼	▾	🟥	🟥	🟥	🟥	🟥
	850	🟩	▲	▲	▲	▲	▴	▴	▴	🟩	▴	▴	▲	▲	▲	▴	▴	▴	🟩	🟩	🟩	▼	▲	▲	▲	▲	▴	🟩	🟩	🟩	🟩	▼	▲	▲	▲	▲	▴	🟩	🟩	🟩	🟩	▼	▼	🟩	▴	▲	▲	▲	▲	▴	▴	▼	▼	🟥	🟩	🟩	🟩	🟩	▴	🟩	🟥
	1000	▲	▲	▲	▲	▲	▴	▴	▴	▴	▴	▲	▲	▲	▲	▲	▴	▴	🟩	🟩	🟩	▴	▲	▲	▲	▲	▴	🟩	🟩	🟩	🟩	▴	▲	▲	▲	▲	▴	🟩	🟥	🟥	🟩	▾	▲	▲	▲	▲	▲	▲	▲	▴	▴	▼	▴	▲	▲	▲	▴	▴	🟩	🟩	🟩
Relative humidity	250	▲	▲	▲	▲	▲	▲	▲	▴	▴	▴	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▴	▴	🟩	🟩	▲	▲	▲	▲	▲	▲	▲	▴	🟩	🟩	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
	700	▲	▲	▲	▲	▲	▲	▲	▴	▴	🟩	▲	▲	▲	▲	▲	▲	▲	▴	▴	▴	▲	▲	▲	▲	▲	▲	▴	▴	🟩	▴	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▼	▼	🟥	▴	▲	▲	▲	▲	▲	▴	▼	🟥	▲	▲	▲	▲	▲	▲	▲	▲
10 m wind speed @ sea		▼	🟩	▴	▲	▴	🟩	▴	▴	🟩	▴	▼	🟩	▴	▲	🟩	🟩	🟩	🟩	🟩	🟩	▼	▾	▲	▲	▲	▲	▴	🟩	🟩	🟩	▼	▾	▴	▲	▴	▴	🟩	🟩	🟥	🟩	▼	▼	🟩	▲	▲	▲	▴	▴	▴	▴	▼	▼	▾	🟥	🟥	🟩	🟥	🟥	🟥	🟥

−∞, ▼, [−3, ▾, −1], 🟥, −0.5], ⬜, [0.5, 🟩, [1, ▴, 3], ▲, ∞.

FIG. 1. A portion of the ECMWF scorecard for IFS cycle 41r2 (implemented 8 Mar 2016) compared to cycle 41r1 (implemented 12 May 2015) for the high-resolution forecasts verified by the respective analyses for 10 Aug 2015–7 Mar 2016. The symbols indicate ranges of the ratio of metric difference to confidence interval width, which is calculated for a paired two-sided t test at $p = 0.05$. The symbols and the ranges (with closed ends indicated by square brackets) are given below the scorecard. [After Fig. 7 of Hólm et al. (2016).]

root-mean-square error (RMSE); however, the discussion is general and applicable to any other similar metrics, both for forecasts and for analyses.

When conducting an impact test comparing alternative or new analysis methods, quality control procedures, forecast model components (e.g., cumulus parameterizations), or observational data sources (e.g., adding new satellite sensors to a NWP data assimilation system), it is often desirable to create summaries of the various assessment metrics. Such summaries can be used to concisely report results and, for better or worse, are sometimes relied on for programmatic decision-making. An example is the ECMWF scorecard that summarizes the impact of recent changes in the Integrated Forecast System (IFS) reported by Hólm et al. (2016). In Fig. 1, which shows a portion of that scorecard, each symbol corresponds to the forecast impact of the system changes on an individual metric. For example, in the rightmost column, which is for RMSE in the tropics, the first line of symbols summarizes the positive (green) impact on 100-hPa temperature, with the most significant results for day 7, while the next line of symbols summarizes the mixed impacts on 250-hPa temperature, with highly significant negative impact on days 1 and 2, followed by both significant and highly significant positive impacts on days 4–10. While a scorecard summarizes a large number of different assessment metrics in just one page, in some situations, a more compact summary may be useful.

Summary assessment metrics (SAMs) can be created if the individual original or primary assessment metrics (PAMs) are first normalized. The normalized assessment metrics (NAMs) are comparable and therefore can be combined. While global SAMs have limited diagnostic usefulness, SAMs created along various dimensions of the NAMs (e.g., as a function of forecast hour) are useful displays of how error varies along such dimensions for different experiments. There are many possible normalizations. One normalization is to convert PAMs into skill scores, for example,

$$S = 1 - r_e^2/r_s^2, \tag{1}$$

where r_e is the experiment RMSE and r_s is the standard or reference RMSE. Both the Met Office (UKMO) NWP index (Rawlins et al. 2007, see their appendix) and the U.S. Air Force (USAF) General Operations (GO) index (Newman et al. 2013; Shao et al. 2016) are defined as $N = 1/\sqrt{1 - S_w}$, where S_w is a weighted sum of the individual skill scores calculated using Eq. (1). The NWP index and the GO index differ in the selection of PAMs, the definition of r_s, and values of the weights. Another normalization is seen in the ECMWF scorecard, where differences between the new and old IFS

PAMs are normalized based on a paired t test for the 95% significance level. Recently, BGK suggested the OFS normalization based on the minimum and maximum in the sample, and applied this to a set of PAMs almost identical to that used here (as described in section 4). Here we propose and explore the use of an alternative normalization based on the probability integral transform (Angus 1994), by making use of the ECDF of a reference sample.

The ECDF normalization for an individual specific PAM is the fraction of the reference sample of similar PAMs worse (i.e., less skillful) than the given individual PAM. This provides a natural way to normalize each PAM into the range [0, 1]. The average of a number of normalized PAMs has a distribution that is asymptotically normal (section 2b). Note that the reference sample can be all similar PAMs from the current experiments, or a sample of similar PAMs from recent operational forecasts.

Advantages of the ECDF approach are that 1) it is nonparametric and hence does not rely on assumptions of normality; 2) it is extremely stable, in particular in comparison to the use of the sample minimum and maximum; 3) allows for flexibility in the choice of reference sample, and therefore can be used for on-going day-by-day monitoring; and 4) can be applied, as is done in section 4, to multiple experiments simultaneously (i.e., does not require a pairwise matchup). As a consequence of the first point the ECDF approach provides more reliable estimates of uncertainty, and therefore confidence intervals.

The calculations of NAMs, SAMs, and the uncertainty of SAMs are detailed in section 2. This calculation is independent of how the reference sample is specified and of the collection of PAMs used. Then in section 3, different approaches to defining the reference sample are considered. The definition of the reference sample should always be included in any report of SAMs. Example results are presented in section 5 for the experiments of BGK, which is briefly described first in section 4. A summary and conclusions are given in section 6.

2. Calculation procedures

The calculation of SAMs involves the following three steps:

- Subset—Define appropriate subsets. For example, one subset could be all initial times for all experiments, for Northern Hemisphere extratropics (NHX) AC for 5-day forecasts of 500-hPa height. Under H_0, the null hypothesis, all the metrics within a subset are from the same reference distribution.
- Normalize—Each PAM is normalized. The resulting NAMs range from 0 (poor) to 1 (excellent). The normalization is different for each subset. ECDF normalization, described below and schematically depicted in Fig. 2, is proportional to rank in the subset of the reference sample. Under H_0, the normalized metrics are uniformly distributed on [0, 1].

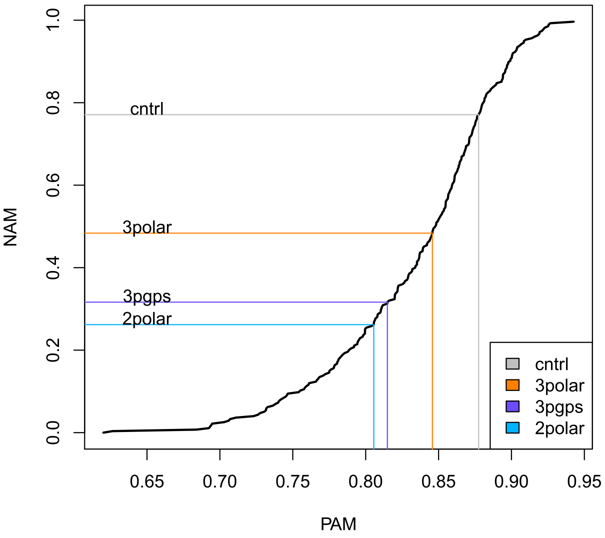


FIG. 2. The ECDF and the transformation from PAMs to NAMs. In this example the ECDF (black curve) is derived from the sample of NHX AC PAMs for 5-day forecasts of 500-hPa height taken from the experiments of BGK. The colored lines show the transformation from PAM to NAM for the forecast initialized at 0000 UTC 18 Jul 2014 for each experiment.

- Average—Since the normalized metrics are comparable, we may average them for each experiment over some or all of the different subset dimensions: variables, levels, forecast times, geographic domains, initial times, and metrics (e.g., AC and RMSE). Under H_0, the averages are approximately Gaussian and have mean 0.5, and variance $1/(12n)$, where n is the number of NAMs averaged.

a. Calculation of normalized assessment metrics (NAMs)

We normalize each PAM value (x) using an ECDF for that metric. Then the NAM is equal to the probability that X, a randomly chosen element from the reference subset distribution is worse than x. We denote this probability by $P(x)$. Stated another way, each NAM (y) is a dimensionless number on the unit interval [0, 1], determined as the quantile of the ECDF corresponding to the PAM value [i.e., $y = P(x) = \Pr(X$ is worse than $x)$]. This discussion applies within each subset (i.e., within the particular variable, level, forecast time, geographic domain, and metric).

Figure 2 illustrates how the normalization works for the subset of NHX AC for 5-day forecasts of 500-hPa height taken from the experiments of BGK. In the figure, the sample of all such forecasts defines the ECDF curve, and the mapping from PAM to NAM is drawn for each forecast initialized at 0000 UTC 18 July 2014. Because of the steepness of the ECDF curve for AC values between 0.80 and 0.90, the 2polar, 3pgps, 3polar, and cntrl experiments' AC values of 0.806, 0.815, 0.846, and 0.877, respectively, are normalized to 0.262, 0.316, 0.484, and 0.771. The use of subsets accounts for the natural variation of forecast skill along the subset dimensions (i.e., from subset to subset). This approach is nonparametric: it does not make any assumption about the distribution of the metrics. Note that we define the NAM to be zero for any x worse than all values in the subset distribution and the NAM to be one for any x better than all values in the subset distribution.

Practically, a NAM is the fraction of cases in the reference subset sample $\mathcal{R}$ that are worse than the particular value of the PAM. Using the rank function results in efficient calculation. First consider a PAM like AC where larger values are better—that is, $x_a < x_b$ implies that case a is worse than case b. If there are n elements in $\mathcal{R}$, and if r is the rank of x in the sample $\mathcal{R} + x$ (i.e., the sample $\mathcal{R}$ with x appended), then the NAM is given by

$$y = \frac{r-1}{n}. \tag{2}$$

The values of y must be in the interval [0, 1] since if x is smaller than all of $\mathcal{R}$, the rank of x in $\mathcal{R} + x$ is 1 and if x is larger than all of $\mathcal{R}$, the rank of x in $\mathcal{R} + x$ is $n + 1$. When calculating the rank, ties are given the minimum possible rank. This method of treating ties—used in golf matches and sometimes called the 1224 rank—has the effect that $(r - 1)$ is always the number of worse (smaller) values. For a metric like RMSE, where smaller values are better, determine the rank of the negative of RMSE, and then apply Eq. (2). For a metric like mean error, where smaller absolute values are better, determine the rank of the negative of the absolute value of the mean error, and then apply Eq. (2). If mean error PAMs are included, then it is preferable to substitute error standard deviation PAMs for RMSE PAMs, since RMSE, being the root-mean-square of error standard deviation and mean error, is not independent of the mean error.

When the reference sample $\mathcal{R}$ is equal to the experiment sample $\mathcal{E}$ composed of all the values of x, then the vector $\mathbf{r}$ of all the values of r can be determined at once by applying a ranking procedure to $\mathcal{R}$.[1] When $\mathcal{R} \neq \mathcal{E}$, $\mathbf{r}$ can be calculated by applying rank to two samples: first to the sample $\mathcal{E}$, then to the sample $\mathcal{E} + \mathcal{R}$ (created by appending $\mathcal{R}$ to $\mathcal{E}$). Then

$$\mathbf{r} - 1 = \mathrm{rank}(\mathcal{E} + \mathcal{R}) - \mathrm{rank}(\mathcal{E}). \tag{3}$$

Consider an example: suppose the rank of x is 5 within $\mathcal{E}$ and 15 within $\mathcal{E} + \mathcal{R}$. Then since x is better than 4 elements in $\mathcal{E}$ and 14 in $\mathcal{E} + \mathcal{R}$, we know that x is better than 10 elements in $\mathcal{R}$.

b. *Summary assessment metrics (SAMs) and significance testing*

Under the null hypothesis (H_0) that there is no impact due to the individual case (initial or valid time) or experiment, each NAM has an independent uniform distribution on the unit interval. This holds exactly if we use the true distribution function of the PAMs (Angus 1994), and the approximation becomes increasingly accurate as the reference sample size increases. Then, each NAM has an expected value of 1/2 and a variance of 1/12. When a set of forecasts is especially good or especially bad, H_0 would be rejected. This allows for significance testing of various combinations of NAMs into SAMs. If a SAM is the average of m NAMs, then that SAM has an expected value of 1/2 and a variance of $1/(12m)$. If we difference two similar SAMs (say for two experiments) then that difference has an expected value of 0 and a variance of $1/(6m)$. In general, under H_0, a SAM has a Bates (1955) distribution. For moderately large m,[2] SAMs will be effectively normally distributed, allowing easy calculation of p values and testing of significance.

3. The ECDF reference sample

The choice of the reference sample for defining the ECDF is critical and will depend on the type of experiment. We outline two alternatives here, but others are possible. In any use of this approach, the reference sample must be clearly defined.

[1] For example, in the R computer language, calculate $\mathbf{r}$ with the command r ← rank(R, ties.method='min', na.last='keep'). Here ties.method='min' chooses 1224 ranking, and na.last='keep' sets the rank of missing values to missing.

[2] In this context, moderately large may be only 10 or so. While uniform random numbers can be transformed into random numbers of any distribution using the inverse probability integral transform (Angus 1994), in the "old" days, 12 random uniform numbers were used to calculate "random normal" numbers. Practical experience over the years shows that 12 is adequate for many purposes, but the choice of 12 versus say 10 or 15 was probably driven by coding considerations to avoid divisions, which can be costly on some current and most early computer architectures. Since the sum of 12 uniform random numbers has a mean of 6 and a variance of 1, the algorithm "add 12 uniform random numbers and subtract 6" produces standard (zero mean, unit variance) random numbers with no division operations.

One sample definition, the self-sample, is the collection of all cases (valid times or initial times) and all experiments. (This is the case $\mathcal{R} = \mathcal{E}$ in section 2.) The self-sample is applicable for impact experiments where we expect fairly large impacts outside the range of an historical sample. By design the average of SAMs over the experiments should be 0.5. Then if there are only two experiments, in plots of SAM versus forecast time or some other subset dimension, the SAM curve for one experiment will be the mirror image about the line $y = 1/2$ of the other experiment. In a comparison of an observing system simulation experiment (OSSE) to an observing system experiment (OSE) using the self-sample for the OSE and a restricted self-sample for the OSSE (restricted to the same experiments conducted in the parallel OSE) could provide a useful tool to calibrate the OSSE to the OSE.

A second reference sample definition, the historical sample, is the collection of all cases from the last year or the last several years close to the same time of year. The historical sample is preferred for preoperational tests of incremental improvements to a forecast system. There are various options for implementing the historical sample: if x is the AC of the 48-h forecast of tropical vector winds at 850 hPa, the reference sample might be the ACs of the 48-h forecast of tropical vector winds at 850 hPa for the previous year for the range $n/2 - 1$ days before to $n/2$ days after the calendar day of the metric we are normalizing. Because this sample definition changes from day to day, the calculation of the rank cannot be optimized as described in section 2. An alternative that would allow some optimization would be to use the three months of the previous year centered on the current month.

Other reference samples could be used. For example, to show the improvement of forecast skill over decades, all the forecasts over that time period could be used to define the ECDF. Using the previous year for each NAM would give a different view—a rate of increase view.

4. Application to a data impact study

Calculations shown below (section 5) are for the OSEs described by BGK. These experiments examine three plausible future data configurations in the global observing system (GOS) that would result in data gaps, and BGK quantify the impacts of these changes in GOS configuration on the skill of the January 2015 NOAA global operational system, which includes the Global Forecast System (GFS) model at T1534 resolution (~13-km horizontal resolution), and the hybrid, ensemble Kalman filter/Gridpoint Statistical Interpolation (GSI) analysis system with 80 ensemble members at T574 resolution (~27-km horizontal resolution), all using 64 vertical levels. The following are the experiment names with brief descriptions.

- cntrl: All satellite and conventional observing systems used in the January 2015 operational implementation are included in this baseline (best case) experiment.
- 3polar: This experiment considers the loss of all secondary and backup polar satellites, retaining only one satellite in each primary (early morning, midmorning, and evening) orbit.
- 3pgps: As in 3polar, but with a decrease in the density of extratropical (poleward of 24°) satellite radio occultation (RO) observations.
- 2polar: As in 3polar, but without the evening platform (i.e., retaining only two polar satellites: one in the early morning orbit and one in the midmorning orbit).

To analyze the results of these experiments, BGK compared a number of statistical metrics (AC, RMSE, and mean error or bias) for several variables (geopotential height, temperature, vector wind, and specific humidity) using different verification datasets (the cntrl analysis, the operational ECMWF analysis, and North American radiosondes). Additional assessment tools compared 6-h quantitative precipitation forecasts to radar/rain gauge precipitation analyses and hurricane track forecasts to best-track estimates. In addition, BGK calculated SAMs using both the OFS and the UKMO NWP index.

In summary, BGK find that "removing secondary satellites results in significant degradation of the forecast. This is unexpected since it is generally assumed that secondary sensors contribute to system's robustness but not necessarily to forecast performance. Second, losing the afternoon orbit on top of losing secondary satellites further degrades forecast performances by a significant margin. Finally, losing extratropical RO observations on top of losing secondary satellites also negatively impacts the forecast performances, but to a lesser degree" (p. 2547). These findings are consistent with the results presented in section 5.

The ECDF assessment metrics for the OSEs of BGK in this paper (here and in section 5) are determined with the self-sample reference sample of section 3 (i.e., all initial times for the four experiments). The PAMs are calculated from the Verification Statistics Database (VSDB; Brill and Iredell 1998) archive[3] for the four

[3] The VSDB files contain the sums needed to calculate both RMSE and AC. For the RMSE calculation, the pres files contain the number of points in the geographical domain, the domain means of the forecast and analysis, and the domain means of the three possible products of the forecast and analysis (i.e., $F \times F$, $F \times A$, $A \times A$). For the horizontal wind vector, the calculations are similar but with vectors replacing scalars, and dot products replacing ordinary multiplications. For the AC calculation, the anom files have the same structure as the pres files, but with the forecast and analysis replaced with the forecast and analysis anomaly with respect to climatology.

experiments using the cntrl analysis for verification. From the VSDB files, the RMSE and AC are calculated for all values of the following:

- variables: for geopotential height, temperature, and vector wind;
- levels: at 250, 500, 700, and 850 hPa;
- forecast times: every 24 h from 1 to 7 days;
- geographic domains: for NHX, Southern Hemisphere extratropics (SHX), and tropics;
- initial times: at 0000 UTC from 25 May until 31 July 2014 (or 0525 to 0731 in mmdd format); and
- experiments: for 2polar, 3pgps, 3polar, and cntrl.

Unlike BGK, we do not include the 0-h forecast time since 0-h errors can be quite different from other forecast times. Also here we use three domains, instead of a single global domain, since forecast skill behavior in the tropics is often quite different than in the extratropics.

Some of these RMSE and AC values are missing because there are no entries in the files archived by BGK. First the AC values are missing for geopotential height at 850 hPa, temperature at 700 hPa, and wind at 700 hPa. Second, all RMSE and AC values are missing for the single 2polar forecast initialized at 0000 UTC 15 July 2014. Note that we include all forecasts times for each initial time. This means that samples for each forecast time are the same size, but samples for valid times at the beginning and end of the experiments vary. For example, 120-h forecasts valid at 0525–0529 and 0806–0807 UTC are not included because they have initial times before 0525 and after 0731 UTC, respectively. For the purpose of plotting some of the figures that follow, missing values are replaced with 0.5 (the expected value under H_0).

5. Example results

Example results are presented here using the ECDF approach for SAMs for the experiments of BGK (as described in section 4). The results presented average over both RMSE and AC metrics because when examined separately they are very similar.

In Fig. 3 all the PAMs ($m > 30\,000$) for one experiment are compressed into a single number. The reference sample in these results and all results that follow is the combination of all initial times for all experiments. Here the NAMs have been averaged over variables, levels, forecast times, geographic domains, initial times, and metrics. In this figure and those that follow, deviations from the expected value (0.5 under H_0) measure the impacts of the different observing system configurations. The larger the deviation, the larger the impact. Positive impacts correspond to increases (and negative impacts correspond to decreases) in forecast accuracy relative to the null hypothesis that the experimental treatments have no effect. Also in this figure and those that follow, confidence intervals are determined as explained in section 2b.

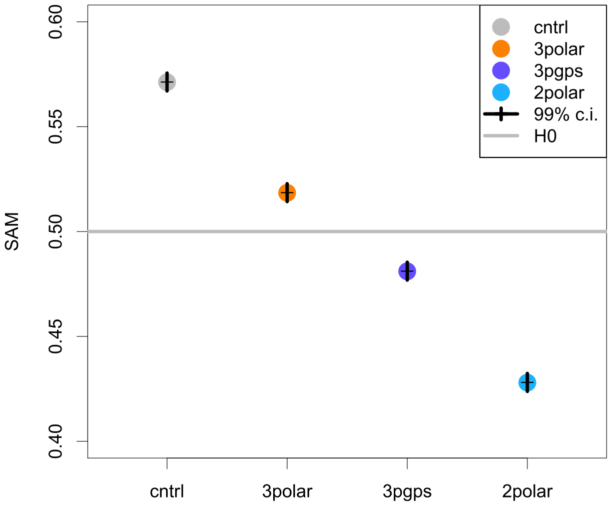


FIG. 3. SAM as a function of experiment alone. The horizontal gray line corresponds to the null hypothesis of no impact. The vertical bar plotted over each colored symbol shows the 99% confidence interval for the result.

The main results obtained by BGK using the OFS are confirmed by Fig. 3:

- The loss of quasi-redundant polar satellite sensors (3polar) results in a significant degradation of overall forecast quality.
- Both removal of the PM polar satellite data and removal of the RO extratropical data lower forecast skill, further degrading forecast quality compared to 3polar.
- Removal of the PM polar satellite data (2polar) has a much larger negative impact than reducing the RO observation coverage (3pgps).

The ECDF uncertainties indicate that these results are statistically robust.

Figure 4 plots SAM as a function of forecast time at different pressure levels and for each experiment. Here the NAMs have been averaged over variables, domains, initial times, and metrics. Impacts range from large positive for cntrl (black) to large negative for 2polar (light blue). All the impacts for these two experiments (i.e., for cntrl and 2polar) are very significant statistically (outside the 0.01–0.99 probability band shown in gray in Fig. 4). Forecast error sources are initial condition errors and model errors. Therefore in these OSEs, where initial condition errors are different, but model errors are similar, there are greater impacts for shorter forecast times. Impacts are somewhat greater for higher levels

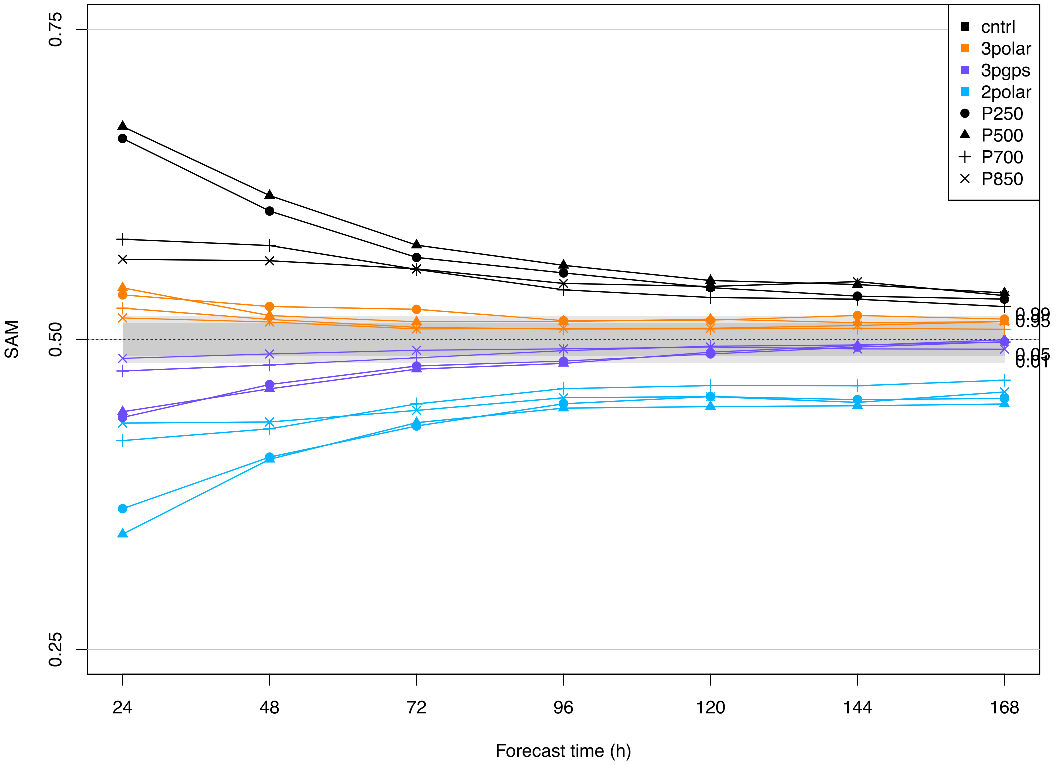


FIG. 4. SAM as a function of forecast time for different levels (symbols) and different experiments (colors).

(250, 500 hPa) and more so at shorter forecast times. Images similar to Fig. 4 may be generated to explore the behavior of forecast skill along other subset dimensions—for example, instead of plotting SAMs for different pressure levels, SAMs might be plotted for different variables or domains.

Figure 5 plots SAM for the different experiments as a function of valid date averaged over forecast days 4, 5, and 6. The ordering of the experiments is fairly consistent: 2polar is usually the worst and cntrl is usually the best. Some dates are more difficult to forecast for all experiments (e.g., 15 June). This phenomenon of exceptional poor forecasts (a.k.a., dropouts) was investigated by Alpert et al. (2009) and Kumar et al. (2009, 2016).

6. Summary and conclusions

The empirical cumulative density function (ECDF) approach can be used to combine multiple, diverse assessment metrics into summary assessment metrics (SAMs) to analyze the results of impact experiments and for preoperational implementation testing with NWP models. The main advantages of the ECDF approach are that it is amenable to statistical significance testing and produces results that are easy to interpret because the SAMs for various subsets tend to vary smoothly and in a consistent manner. In addition, the ECDF approach can be applied in various contexts thanks to the flexibility allowed in the definition of the reference sample (section 3).

SAMs could be weighted averages of NAMs. In this paper SAMs are simple averages of NAMs, but there are several reasons to use a weighted average. First, there may be a desire to weight some components more highly because of their relevance to a particular forecasting situation. For example, PAMs related to upper-level winds would be of greater interest for aviation forecasting. Second, some of the PAMs might be correlated. For example, if four PAMs were strongly correlated, rather than eliminate three of these and having to decide which to keep, the associated weights might be reduced. Third, some PAMs might be particularly sensitive to exceptional poor forecasts (i.e., dropouts; Kumar et al. 2009, 2016) and it might be desirable to increase the associated weights. If this is not done, the ECDF approach will tend to hide the dropout signal—by design the normalization eliminates differences in distributions. Very large differences in PAMs that have long tailed distributions become homogenized and are no longer exceptional once normalized. It should be noted that dropout cases are worthy of synoptic evaluation

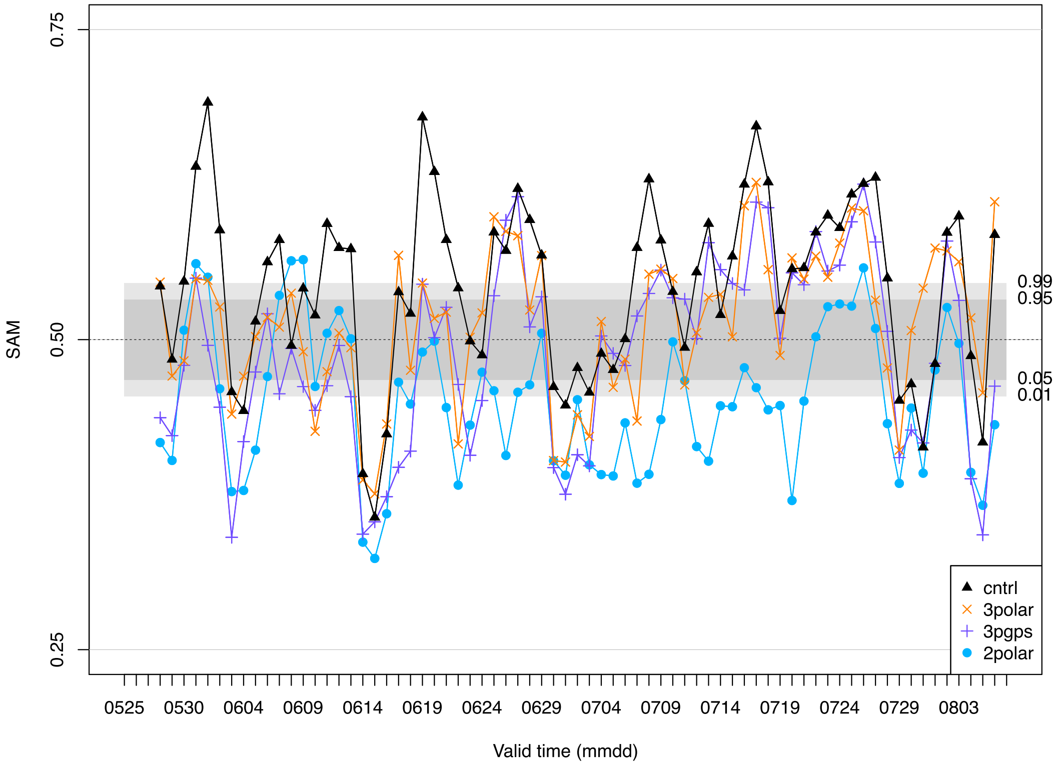


FIG. 5. SAM as a function of valid time (mmdd format) for different experiments (colors and symbols). These results are for SAMs averaged over the 96-, 120-, and 144-h forecast times.

that goes well beyond the statistical assessments that are the subject of this paper. For weighted averaged SAMs, the presentation in section 2b is unchanged except for that the estimate of variance of the SAM becomes $(\sum_i w_i^2)/[12(\sum_i w_i)^2]$, where w_i is the weight associated with the ith NAM. It should be acknowledged that choosing weights, or indeed selecting which PAMs to include,[4] is often ad hoc and introduces a degree of subjectivity in otherwise objective assessments.

The main (perhaps only) assumption of the ECDF approach is the null hypothesis that all the members of a subset are from the same distribution. For OSEs or OSSEs, the subset will usually include all initial (or valid) times for all experiments. When we reject the null hypothesis, we would like to attribute the impact to the differences between experiments, and this is reasonable when the SAMs under consideration include all valid times. However, as seen in Fig. 5 there can be variations with valid time. It might be possible to reduce such variations by applying the ECDF method as described here to differences of PAMs at the same valid time—in the present case these might be 2polar–cntrl, 3pgps–cntrl, and 3polar–cntrl. In such an application, it would be consistent with the null hypothesis of no impact due to the observing systems configuration to choose the reference sample as the differences of PAMs from all possible experiment pairs. In the present case there would be 12 such pairs, including for example, both 2polar–3polar and 3polar–2polar.

The interpretations of the examples presented here (section 5) are consistent with the previously reported conclusions of BGK and with some but not all conclusions of other data gap studies (Cucurull and Anthes 2015; Lord et al. 2016). An interesting finding is that the impact of observations decreases with increasing forecast time. We expect differences in initial conditions to grow. However, it is likely that NWP model error grows more quickly since model error is added at every time step. Further, in these experiments, model error is similar from experiment to experiment since the same model is used in each experiment. Therefore, NWP model error is expected to tend to mask the impact of the differences in initial conditions with increasing forecast time. Also, as seen in Fig. 4, out to 72 h, the impacts are greater higher in the atmosphere, possibly because the data assimilation system extracts more information there. There are two potential contributing factors: first, there are more higher-peaking satellite

[4] Since selection is the ultimate 0 or 1 weighting scheme.

radiance channel observations that pass the cloud quality control; and second, the data assimilation system makes use of channel subsets in which higher-peaking channels have been preferentially selected since less information can currently be extracted from channels sensitive to the boundary layer and lower troposphere because representativeness errors are greater for such channels.

Acknowledgments. Fanglin Yang patiently explained the intricacies of the VSDB file structure. ECMWF gave the authors permission to use the image in Fig. 1. Financial support for this work is gratefully acknowledged, including funding provided by the Disaster Relief Appropriations Act of 2013 (H.R. 152) and NOAA Grant NA14OAR4830103.

REFERENCES

Alpert, J. C., D. L. Carlis, B. A. Ballish, and V. K. Kumar, 2009: Using ''pseudo'' RAOB observations to study GFS skill score dropouts. *23rd Conf. on Weather Analysis and Forecasting/19th Conf. on Numerical Weather Prediction*, Omaha, NE, Amer. Meteor. Soc., 5A.6. [Available online at https://ams.confex.com/ams/23WAF19NWP/techprogram/paper_154268.htm.]

Angus, J. E., 1994: The probability integral transform and related results. *SIAM Rev.*, **3**, 652–654, doi:10.1137/1036146.

Bates, G. E., 1955: Joint distributions of time intervals for the occurrence of successive accidents in a generalized Polya scheme. *Ann. Math. Stat.*, **26**, 705–720, doi:10.1214/aoms/1177728429.

Boukabara, S.-A., K. Garrett, and V. K. Kumar, 2016: Potential gaps in the satellite observing system coverage: Assessment of impact on NOAA's numerical weather prediction overall skills. *Mon. Wea. Rev.*, **144**, 2547–2563, doi:10.1175/MWR-D-16-0013.1.

Brill, K. F., and M. D. Iredell, 1998: EMC verification database. NCEP, Camp Springs, MD, 8 pp. [Available online at http://www.emc.ncep.noaa.gov/mmb/papers/brill/VSDBformat.txt.]

Cucurull, L., and R. A. Anthes, 2015: Impact of loss of U.S. microwave and radio occultation observations in operational numerical weather prediction in support of the U.S. data gap mitigation activities. *Wea. Forecasting*, **30**, 255–269, doi:10.1175/WAF-D-14-00077.1.

Hólm, E., R. Forbes, S. Lang, L. Magnusson, and S. Malardel, 2016: New model cycle brings higher resolution. *ECMWF Newsletter*, No. 147, ECMWF, Reading, United Kingdom, 14–19. [Available online at http://www.ecmwf.int/en/elibrary/16299-newsletter-no147-spring-2016.]

Kumar, V. K., J. C. Alpert, D. L. Carlis, and B. A. Ballish, 2009: Investigation of NCEP GFS model forecast skill ''dropout'' characteristics using the EBI index. *23rd Conf. on Weather Analysis and Forecasting/19th Conf. on Numerical Weather Prediction*, Omaha, NE, Amer. Meteor. Soc., 13A.1. [Available online at https://ams.confex.com/ams/23WAF19NWP/techprogram/paper_154282.htm.]

——, E. Maddy, J. C. Alpert, and S. A. Boukabara, 2016: Global forecast dropout prediction tool in support of the NCEP model evaluation group (MEG)—A collaborative project between JCSDA/NESDIS & NWS. *20th Conf. on Integrated Observing and Assimilation Systems for the Atmosphere, Oceans, and Land Surface (IOAS-AOLS)*, New Orleans, LA, Amer. Meteor. Soc., 6.3. [Available online at https://ams.confex.com/ams/96Annual/webprogram/Paper288630.html.]

Lord, S., G. Gayno, and F. Yang, 2016: Analysis of an observing system experiment for the Joint Polar Satellite System. *Bull. Amer. Meteor. Soc.*, **97**, 1409–1425, doi:10.1175/BAMS-D-14-00207.1.

Newman, K. M., Zhou, C., M. Hu, and H. Shao, 2013: Configuration testing of GSI within an operational environment. *17th Conf. on Integrated Observing and Assimilation Systems for the Atmosphere, Oceans, and Land Surface (IOAS-AOLS)*, Austin, TX, Amer. Meteor. Soc., 620. [Available online at https://ams.confex.com/ams/93Annual/webprogram/Paper221922.html.]

Rawlins, F., S. P. Ballard, K. J. Bovis, A. M. Clayton, D. Li, G. W. Inverarity, A. C. Lorenc, and T. J. Payne, 2007: The Met Office global four-dimensional variational data assimilation scheme. *Quart. J. Roy. Meteor. Soc.*, **133**, 347–362, doi:10.1002/qj.32.

Shao, H., and Coauthors, 2016: Bridging research to operations transitions: Status and plans of community GSI. *Bull. Amer. Meteor. Soc.*, **97**, 1427–1440, doi:10.1175/BAMS-D-13-00245.1.