P618 A Non-Parametric Definition of Summary NWP Forecast Assessment Metrics: Application to Polar Data Gap Impact Assessment and NWP Centers Skills Inter-Comparison*

Ross N. Hoffman†
NOAA Atlantic Oceanographic and Meteorological Laboratory, Miami Florida &
Cooperative Institute for Marine and Atmospheric Studies,
University of Miami, Miami, Florida
Sid-Ahmed Boukabara
NOAA/NESDIS/Center for Satellite Applications and Research (STAR),
College Park, Maryland
Krishna Kumar
Riverside Technology Inc., at NOAA/NESDIS/Center for Satellite Applications
and Research (STAR), College Park, Maryland
Kevin Garrett
Riverside Technology Inc., at NOAA/NESDIS/Center for Satellite Applications
and Research (STAR), College Park, Maryland
Sean P. F. Casey
NOAA Atlantic Oceanographic and Meteorological Laboratory, Miami, Florida &
Cooperative Institute for Marine and Atmospheric Studies,
University of Miami, Miami, Florida
Robert Atlas
NOAA Atlantic Oceanographic and Meteorological Laboratory, Miami, Florida

ABSTRACT

Summary assessment metrics (SAMs) are defined as the average of a collection of normalized assessment metrics (NAMs). A normalization based on the empirical *c.d.f.* (ecdf) is proposed and tested for two cases. For the OSE case, results are consistent with the conclusions of Boukabara et al. (2016). For the NWP centers case, results generally agree with our prior assessment of relative forecast skill.

1. Introduction

Standard NWP verification systems generate a multitude of forecast skill metrics, which we will call primary assessment metrics (PAMs). Typically, anomaly correlation (AC) and root mean square error (RMSE) statistics are PAMs that might be calculated along several dimensions such as domain, vertical level, variable, forecast length and either verification time or initial time, all for a number of treatments. For example the domains might include the

*Extended abstract for the poster presentation of Hoffman et al. (2017a).

†ross.n.hoffman@noaa.gov

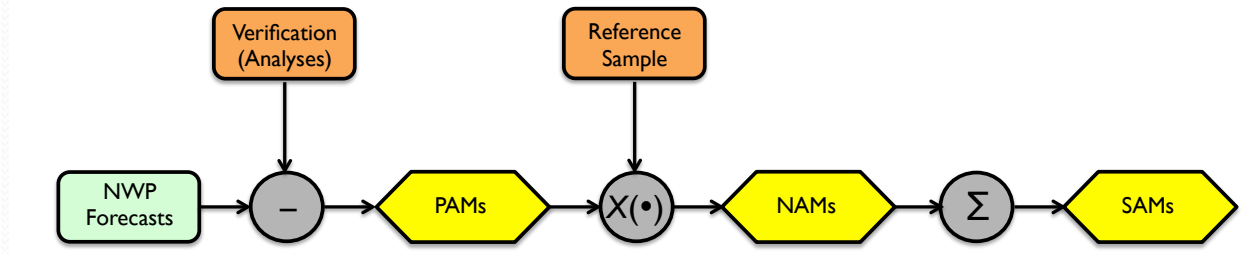

FIG. 1: Flow chart showing the procedure of calculating PAMs, NAMs, and SAMs. Assessment metrics are shown as yellow hexagons, processes as grey circles, fixed data sets as orange rounded rectangles, and the forecast data set as the green rounded rectangle.

Northern Hemisphere extratropics (NHX), the tropics (TRO), and the Southern Hemisphere extratropics (SHX). The treatments might be different NWP centers or different experiments (different observing systems, different model configurations, etc.). Typically, we want to compare the different metrics across treatments. In some situations it is desirable to summarize these metrics and/or to compare the metrics along one or two of the dimensions listed just above.

Hoffman et al. (2017b, hereafter HBK) describe the empirical *c.d.f.* (ecdf) approach to normalize PAMs into normalized assessment metrics (NAMs) so that they may be combined into summary assessment metrics (SAMs). Figure 1 shows the process flow from forecasts and verifications to PAMs, NAMs, and SAMs. Details are given in Section 2. HBK applied the ecdf approach to the OSE study of Boukabara et al. (2016, hereafter BGK). In Section 3, we summarize HBK and present some additional results for the OSE study. In addition the ecdf SAMs were calculated for 6 NWP centers for 3 months in 2015. These calculations and results are described in Section 4. A summary and plans for future work are given in the concluding remarks (Section 5).

2. Methodology

Once PAMs are calculated, there are three principal steps to calculate NAMs and SAMs (Fig. 1). These are to define the reference sample, normalize the PAMs to create the NAMs, and average the NAMs into SAMs. We summarize these three steps here. See HBK for details.

a. *The reference sample*

An example of a reference sample for an OSE or OSSE is all experiments, all initial times, for NHX AC for 5-day forecasts of 500 hPa height. Under H_0, the null hypothesis, all the members of a subset are from the same distribution.

The choice of the reference sample for defining the ecdf is critical and will depend on the type of experiment. In any use of this approach the sample must be clearly defined. Types of reference samples include the self-sample and the historical sample. The self-sample, is the collection of all cases (valid times or initial times) and all experiments. Consequently, the average of NAMs and SAMs over the experiments and cases is 0.5. All results presented here use the self-sample. 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 pre-operational tests of incremental improvements to data assimilation (DA) and forecast systems.

b. *Normalization*

Each PAM is converted to a NAM that ranges from 0 (poor) to 1 (excellent). The normalization depends on the subset. The ecdf normalization is proportional to rank in the reference sample. Under H_0, the NAMs are uniform on [0,1]. The ecdf score for a PAM is the fraction of cases in the sample, for which this particular PAM is better. Figure 2 shows the ecdf for 5-day forecasts of NHX 500 hPa height AC for the OSEs of BGK, and the normalization for the particular forecasts initialized 00 UTC 18 July 2014.

Some previous SAMs include the UKMO NWP in-

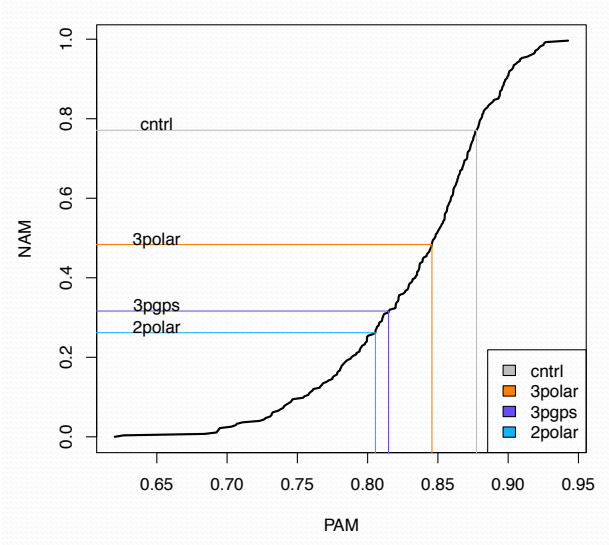

FIG. 2: The ecdf (black curve) and the transformation from PAMs to NAMs (colored lines) for the forecast initialized 00 UTC 18 July 2014 for each experiment. (After Fig. 2 of HBK.)

dex and the USAF General Operations (GO) index—both based on weighted sums of (normalized) skill scores—and the Overall Forecast Score (OFS) of BGK, which is based on a min-max normalization. Additional normalizations have been employed in so-called scorecards where each NAM is plotted graphically, usually in a single tabular array. A scorecard example is given in HBK, which is an extract of the ECMWF scorecard comparing IFS cycle 41r2 (Mar 2016) to IFS cycle 41r1 (May 2015) presented by Hólm et al. (2016). In the ECMWF scorecard the normalization is based on a t-test confidence interval. See HBK for more details.

c. *Averaging*

Since the NAMs are comparable, we may average them over dimensions and values. Under H_0, the averages (SAMs) are approximately Gaussian with mean 1/2, and variance $1/(12n)$. Here n is the number of NAMs if they are independent or an effective number of NAMs if they are correlated.

In the figures of SAMs, deviations from the expected value of 0.5 measure the impacts of the different observing system configurations. The larger the deviation, the larger is the impact. Positive (negative) impacts correspond to increases (decreases) in forecast accuracy relative to the null hypothesis that the experimental treatments have no effect. In the figures, approximate 5–95% and 1–99% confidence intervals are plotted for the null hypothesis. These confidence intervals are calculated assuming that the individual NAMs are independent. This is not true in practice, but estimating the effective number of NAMs is difficult.

SAMs might be defined as weighted averages of NAMs. We have adopted a simple approach of using all PAMs that are usually assessed individually for calculating NAMs and SAMs with unit weighting. The unweighted approach has the advantage of being both simple and fair, with no attempt to subjectively adjust weights. Of course the individual NAMs are correlated potentially along most dimensions including forecast length, vertical level, etc. For example, typical decorrelation intervals for initial times and forecast times are 12 and 18 h respectively. We expect that there are correlations to different degrees between practically all metrics generated. These correlations are in a sense inherent to the creating SAMs since the same DA and forecast systems are the source of an array of PAMs. Correlations between individual NAMs have a relation to using weights to calculate the SAMs. In the case with no weights, any correlations have the effect of giving additional weight than one should to the correlated NAMs and lower the effective number of NAMs that determine the uncertainty of the SAMs.

3. An OSE example

The experimental setup of BGK makes use of the January 2015 NOAA global operational system, which includes Global Forecast System (GFS) model at T1534L64 resolution and the hybrid, ensemble Kalman filter/GSI analysis system with 80 ensemble members at T574L64 resolution. Four observing system configurations are included in the experiments of BGK:

- cntrl: All observing systems used in operations. This is the best-case experiment.
- 3polar: Retains only one satellite in each primary orbit (early-AM, mid-AM, PM).
- 3pgps: Like 3polar, but with greatly reduced RO observations poleward of 24°.
- 2polar: Like 3polar but without the PM satellite.

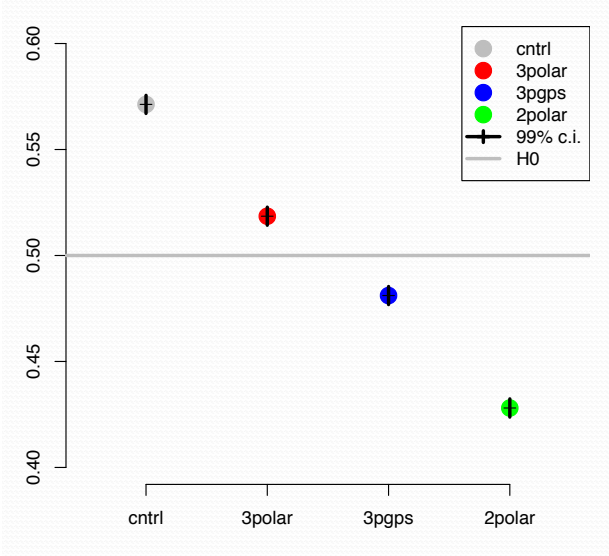

FIG. 3: SAM as a function of experiment alone for the OSEs. (After Fig. 3 of HBK.)

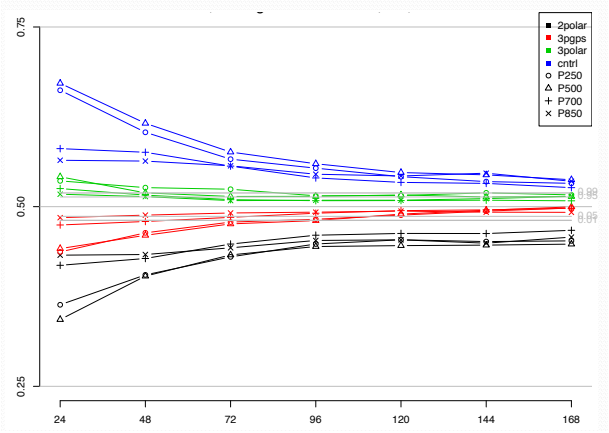

FIG. 4: SAM as a function of forecast time for different levels (symbols) and different experiments (colors) for the OSEs. (After Fig. 4 of HBK.)

The reference sample is all initial times for all four experiments. In calculating the PAMs, the cntrl analysis is used for verification for all four experiments. For the OSE example, the PAM dimensions and coordinate values along these dimensions are:

- variables :: geopotential height (HGT), temperature (T), and vector wind (WIND).
- levels :: 250, 500, 700, 850 hPa.
- forecast times :: every 24 hours from 1 to 7 days.
- geographic domains :: NHX, TRO, SHX.
- initial times :: 00 UTC from 25 May until 31 July 2014.

For context, we reproduce two figures from HBK. Figure 3 shows the overall SAMs for the BGK OSEs, which confirm the BGK findings that

$$\text{cntrl} > \text{3polar} > \text{3pgps} > \text{2polar}.$$

In this example all the scores ($n > 30,000$) for one experiment are condensed into a single number. In the figure, the grey line corresponds to the null hypothesis of no impact. The vertical bar over plotted each colored symbol shows the 99% confidence interval for the result.

Figure 4 plots SAM as a function of forecast time for different levels and different experiments. In general, there are greater impacts for shorter forecast times. In this OSE examples where initial condition (IC) errors are different, but model errors are similar, NWP model error is expected to become dominant and to tend to mask the impact of the differences in ICs with increasing forecast time. In Fig. 4, we also see (out to 72 h) there are greater impacts higher in the atmosphere, possibly because the data assimilation system extracts more information there.

The next two figures are similar to Fig. 4, but plot SAMs for different domains and different variables instead of for different vertical levels. In Fig. 5, impacts are greatest in the SHX and least in the tropics. This trend is emphasized at short forecast times. In the SHX, conventional data are not plentiful enough to moderate the impact of changes in the satellite data coverage. In the tropics, forecast skill, and hence the potential for impact, is generally low at longer forecast times. Note that longer-range forecasts in the SHX for 2polar are particularly poor compared to those in the NHX. The situation for 2polar in the SHX is very data poor. In Fig. 6, wind impacts are greater than mass field impacts. This is most evident at short forecast times. This seems counter-intuitive since satellite data are more directly related to temperature and hence to geopotential height than wind. It may be an effect of scale, since there is greater wind variability at smaller scales compared to the mass field variables.

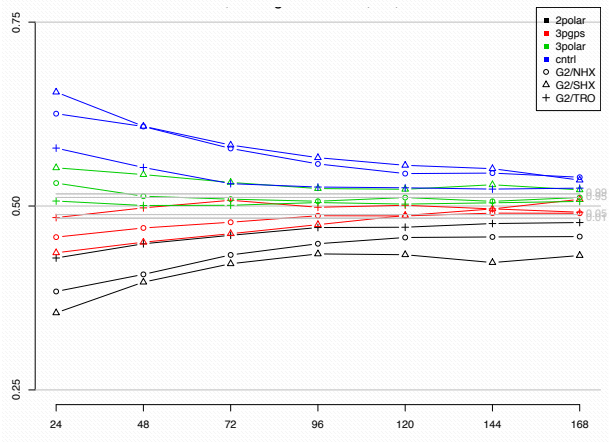

FIG. 5: SAM as a function of forecast time for different domains (symbols) and different experiments (colors) for the OSEs.

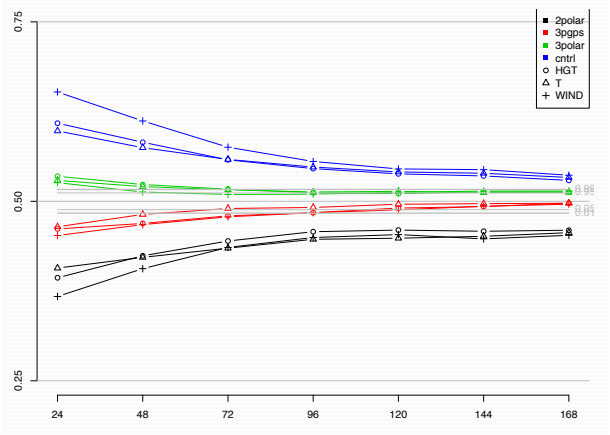

FIG. 6: SAM as a function of forecast time for different variables (symbols) and different experiments (colors) for the OSEs.

4. NWP centers SAMs

NCEP collects forecasts of other NWP centers and routinely calculates PAMs. Here, we use the Canadian (CMC), European (ECM), U.S. Navy (FNO), NCEP (GFS), Indian (NCMRWF), and United Kingdom (UKM) NWP centers PAMs. Each center's analysis is used for verification. We retrieved the archived VSDB data sets for 2015. These contain the sums needed to calculate various PAMs, including AC and RMSE. For this study, the reference sample is all initial times for all centers, month by month and the PAM dimensions and coordinate values along these dimensions are:

- variables :: HGT, T, WIND.
- levels :: 250, 500, 700, 850, 1000 hPa.
- forecast times :: every 24 hours from 1 to 6 days.
- geographic domains :: NHX, TRO, SHX.
- valid times :: 00 UTC from 01 until 31 of each month in 2015.

Figure 7 shows the combined SAMs by month for different NWP centers. Overall

$$\text{ECM} > \text{GFS} \sim \text{UKM} > \text{CMC} \sim \text{FNO} > \text{NCMRWF}.$$

Clearly the month-to-month variation is small. Note that for combining AC and RMSE SAMs, a weight of

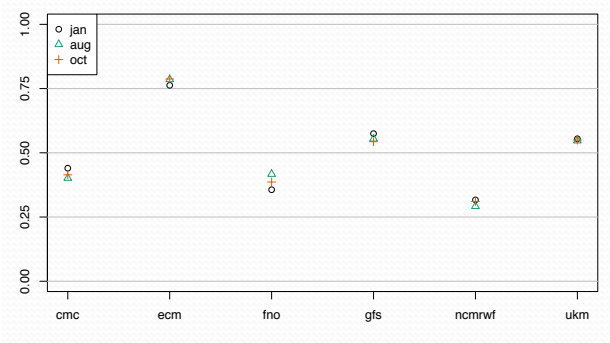

FIG. 7: The combined SAMs by month (symbols and colors) for different NWP centers.

2/5 is given to the AC SAM and 3/5 to the RMSE SAM. This is done because for AC we only have VSDB data for 10 of the 15 possible variable-level combinations. (By default, there are no VSDB entries to calculate AC for HGT(P850), T(P700,P1000), and WIND(P700,P1000).)

The remaining figures are for January only, and with one exception for RMSE SAMs only. The uncertainty bounds presented are consistent with a null hypothesis that all the NAMs are independent draws from a uniform (0,1) distribution. To put bounds on the effective sample size, consider that Fig. 7 shows that the month-to-month variation of overall ecdf SAMs is on the order of 0.025. This puts an upper bound on the sampling uncertainty of the individual monthly values, since the uncertainty in Fig. 7 includes both sampling error and real month-to-month

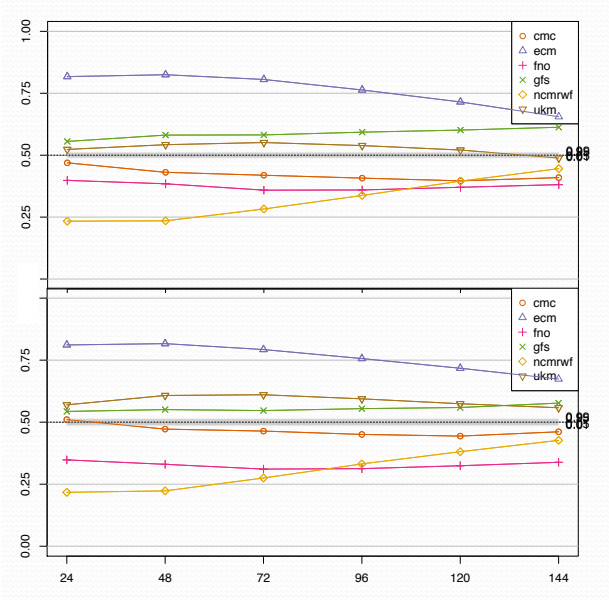

FIG. 8: January RMSE (top) and AC (bottom) SAMs as a function of forecast time for different NWP centers.

variation. The formal 1-σ uncertainty (not plotted) is order of 0.0025 (for n=13950). If we assume all the uncertainty in Fig. 7 is due to sampling uncertainty, then we should increase the formal uncertainty by a factor of 10, or equivalently, we should reduce the effective sample size by a factor of 100. This adjustment would be smaller for subsets. For example, SAMs for a single vertical level, require no adjustments for vertical correlations.

Figure 8 shows January RMSE and AC SAMs as functions of forecast time for different NWP centers.

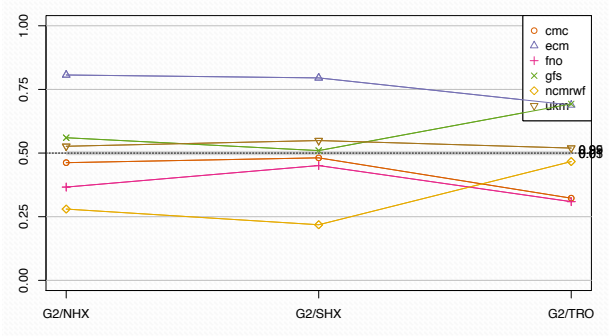

FIG. 9: January RMSE SAMs versus domain for different NWP centers.

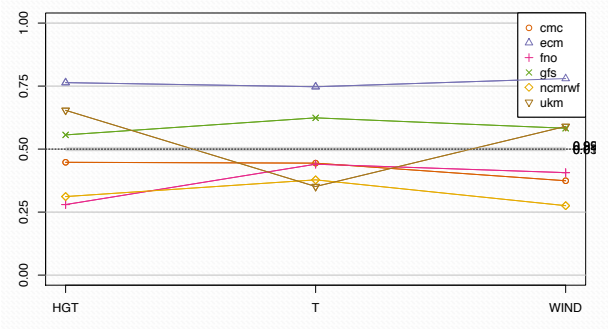

FIG. 10: January RMSE SAMs versus variable for different NWP centers.

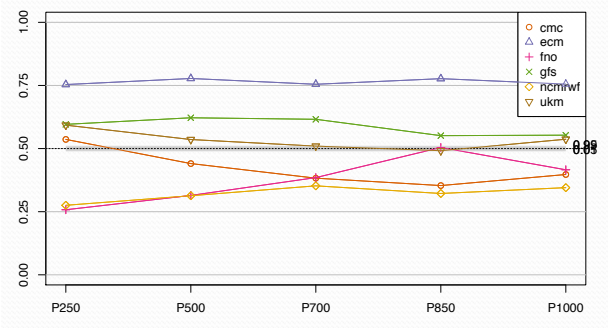

FIG. 11: January RMSE SAMs versus pressure level for different NWP centers.

There is a tendency for the extreme score to tighten up with time. For example, ECM drops from 0.8 to 0.55 and NCMRWF increases from 0.25 to 0.45 as the effect of IC errors lessens. Models that improve relative to others over time likely have smaller model errors (GFS, NCMRWF). Comparing AC to RMSE, UKM and CMC do better in terms of AC, FNO does worse. Figure 9 shows January RMSE SAMs versus domain for different NWP centers. GFS is quite good in the tropics, but not in the SHX, where UKM is better. Figure 10 shows January RMSE SAMs versus variable for different NWP centers. Note that UKM is poor for T compared to HGT and WIND. Figure 11 shows January RMSE SAMs versus pressure level for different NWP centers. Note that FNO does considerably better in the lower atmosphere, especially at 850 hPa than at higher levels. Figure 12 shows January AC SAM versus valid time for different NWP centers. Note that there is a general dropout in forecast skill around 20 January 2015 for all centers.

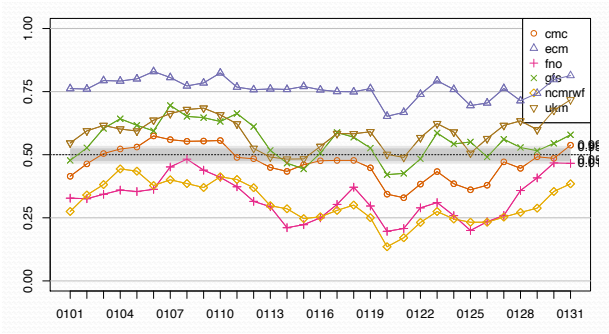

FIG. 12: January RMSE SAMs versus valid time for different NWP centers.

5. Concluding remarks

a. *Summary*

Summary assessment metrics (SAMs) are defined as the average of a collection of normalized assessment metrics (NAMs). A normalization based on the empirical *c.d.f.* (ecdf) is proposed and tested for two cases. An advantage of the ecdf approach is that it is amenable to statistical significance testing. The ecdf SAMs are relatively easy to interpret since the metrics for various subsets vary relatively consistently. For the OSE case, results are consistent with the conclusions of BGK. For the NWP centers case, results generally agree with our prior assessment of relative forecast skill. There are some interesting detailed findings for the NWP centers case. For example, UKM does poorly at forecasting temperature compared to geopotential and wind and FNO is best at forecasting at 850 hPa.

b. *Future work*

We plan to extend our work with the NWP centers PAMs. First, to treat all of 2015, we will experiment with different ways to account for annual cycle in skill, including putting everything into one reference sample, using a moving window reference sample (e.g., use the three months centered on a given month), and applying the SAM calculation to skill differences. Second, we would like to extend this analysis to 2016 and beyond. Third, we will apply the ecdf normalization to other metrics, such as the absolute value of the mean error (AME). Fourth, we will consider the possibility of extending the work to weighted SAMs. Fifth, we will develop approaches to estimate the effective sample size in order to refine our estimates of uncertainty. See the conclusions of HBK for additional discussion of some of these future directions.

Acknowledgement Financial support for this work is gratefully acknowledged, including funding provided by the Disaster Relief Appropriations Act of 2013 (H.R. 152).

References

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.

Hoffman, R. N., S. A. Boukabara, V. K. Kumar, K. Garrett, S. Casey, and R. Atlas, 2017a: A non-parametric definition of summary NWP forecast assessment metrics. *28th Conference on Weather Analysis and Forecasting/24th Conference on Numerical Weather Prediction*, American Meteorological Society, Boston, MA, Seattle, Washington, poster 618. Available online at https://ams.confex.com/ams/97Annual/webprogram/Paper309748.html.

Hoffman, R. N., S.-A. Boukabara, V. K. Kumar, K. Garrett, S. P. F. Casey, and R. Atlas, 2017b: An empirical cumulative density function approach to defining summary NWP forecast assessment metrics. *Mon. Wea. Rev.*, in press. doi:10.1175/MWR-D-16-0271.1.

Hólm, E., R. Forbes, S. Lang, L. Magnusson, and S. Malardel, 2016: New model cycle brings higher resolution. *ECMWF Newsletter*, Spring (147), 14–19.