Progress in Forecast Skill at Three Leading Global Operational NWP Centers during 2015–17 as Seen in Summary Assessment Metrics (SAMs)

ROSS N. HOFFMAN,[a,b] V. KRISHNA KUMAR,[c,d] SID-AHMED BOUKABARA,[d] KAYO IDE,[e] FANGLIN YANG,[f] AND ROBERT ATLAS[a]

[a] *NOAA/Atlantic Oceanographic and Meteorological Laboratory, Miami, Florida*
[b] *Cooperative Institute for Marine and Atmospheric Studies, University of Miami, Miami, Florida*
[c] *Riverside Technology Inc., College Park, Maryland*
[d] *NOAA/NESDIS/STAR, College Park, Maryland*
[e] *University of Maryland, College Park, College Park, Maryland*
[f] *NOAA/NCEP/Environmental Modeling Center, College Park, Maryland*

(Manuscript received 13 July 2018, in final form 17 September 2018)

ABSTRACT

The summary assessment metric (SAM) method is applied to an array of primary assessment metrics (PAMs) for the deterministic forecasts of three leading numerical weather prediction (NWP) centers for the years 2015–17. The PAMs include anomaly correlation, RMSE, and absolute mean error (i.e., the absolute value of bias) for different forecast times, vertical levels, geographic domains, and variables. SAMs indicate that in terms of forecast skill ECMWF is better than NCEP, which is better than but approximately the same as UKMO. The use of SAMs allows a number of interesting features of the evolution of forecast skill to be observed. All three centers improve over the 3-yr period. NCEP short-term forecast skill substantially increases during the period. Quantitatively, the effect of the 11 May 2016 NCEP upgrade to the four-dimensional ensemble variational data assimilation (4DEnVar) system is a 7.37% increase in the probability of improved skill relative to a randomly chosen forecast metric from 2015 to 2017. This is the largest SAM impact during the study period. However, the observed impacts are within the context of slowly improving forecast skill for operational global NWP as compared to earlier years. Clearly, the systems lagging ECMWF can improve, and there is evidence from SAMs in addition to the 4DEnVar example that improvements in forecast and data assimilation systems are still leading to forecast skill improvements.

1. Introduction

Since the start of numerical weather prediction (NWP), a few key statistics have been used to summarize the skill of operational forecasts. With early models, such as the equivalent barotropic model, which have essentially one level and one variable, a single RMSE[1] statistic at a few forecast times was sufficient. So-called headline scores like the 120-h 500-hPa Northern Hemisphere extratropics (NHX) geopotential height anomaly correlation (AC) continue this custom and tend to dominate discussions of forecast skill. We call such an individual unnormalized skill score a primary assessment metric (PAM).

Figure 1, from the WMO Lead Centre for Deterministic NWP Verification (LCDNV) website (http://apps.ecmwf.int/wmolcdnv/scores/time_series/500_z), shows the evolution of the closely related 120-h 500-hPa NHX geopotential height RMSE PAM over the last two decades. It is clear in Fig. 1 that this particular PAM is improving, but improvements are increasingly difficult to achieve, and currently are nil or nearly so for many of the models. This may be because the NWP methods have advanced over the years, and practical predictability is approaching the intrinsic predictability limit (see Lorenz 1982). However, even as some centers show no apparent improvements in skill for this PAM, it is clear that ECMWF continues to improve and thus there is hope for improvements for the other centers. Similar results are seen for other PAMs over this 20-yr time period.

[1] All acronyms are defined in section e of the appendix.

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

DOI: 10.1175/WAF-D-18-0117.1

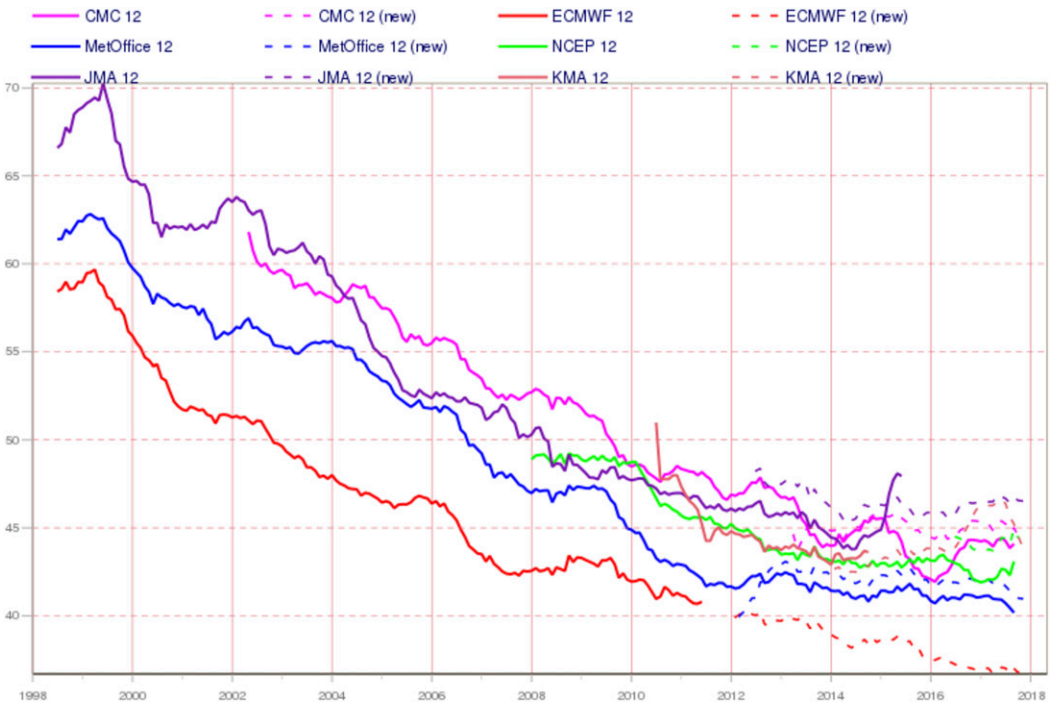

FIG. 1. The evolution of the annually averaged 120-h 500-hPa NHX geopotential height RMSE over the last two decades for the centers indicated in the legend. (Source: http://apps.ecmwf.int/wmolcdnv/scores/time_series/500_z.)

The first goal of the present study is to document the evolution of forecast skill for three of the leading forecast centers—ECMWF, NCEP, and UKMO—over the last 3 years (2015–17). The major upgrades at these centers during this period are listed in Table 1. Daily 1–7-day forecasts initialized at 0000 UTC were verified against the analysis (the 0-day forecast if you will). Each center was verified against itself. A large array of statistical quantities was then available for analysis. The second goal of this study is to present a technique called the summary assessment metric (SAM) method to combine and examine this array of statistics.

Over time, forecast centers have increased the complexity of their atmospheric models, advanced data assimilation methods and strategies, developed and implemented ensemble forecasting techniques, and coupled their atmospheric models to models of other earth system components. Consequently, a large number of PAMs are available for assessing the impact of a change in the global observation system, a change in the use of the observations, or a change in a model. This has led to increasingly complex scorecards that attempt to normalize and display a large number of the PAMs in order to provide a comprehensive summary as an aid to implementation decision-making. For example, scorecards supporting the implementation of ECMWF IFS Cycle 45r1 are given in Figs. 1 and 2 of Buizza et al. (2018). For normalization of the PAMs, ECMWF calculates a paired Student's t statistic in which autocorrelations are accounted for by modeling the time series of paired differences as AR(1)[2] processes (Geer 2016). We call such individual normalized skill scores normalized assessment metrics (NAMs).

Naturally, numerous ways to combine a scorecard or some other collection of NAMs have been proposed. We use the term summary assessment metric to denote such a combination of NAMs. For example, the UKMO NWP index (Rawlins et al. 2007; see their appendix) and the USAF General Operations (GO) index (Newman et al. 2013; Shao et al. 2016) are SAMs that are defined in terms of a weighted sum of NAMs. In these cases the PAMs are all RMSEs and each is divided by a tabulated reference RMSE in producing the corresponding NAM. The choice of weights and the reference RMSEs allow these SAMs to be tailored for individual customers. Boukabara et al. (2016) introduced the overall forecast score (OFS) as a SAM based on a simple formulation of using the minimum and maximum value of a reference sample of PAMs to scale each PAM value to be within

[2] Here, AR(p) indicates autoregressive, order p.

TABLE 1. Major upgrades reported by the three centers during the study period (2015–17). The index i is plotted in a number of the figures at the date of the upgrade. The upgrades are identified here only by version number and/or keywords. (All acronyms are defined in section e of the appendix.) The column labeled Δ (%) is the change in SAM in Fig. 12 due to the upgrade. For details refer to the sources cited.

Center	i	Date	Upgrade	Δ	Source
ECMWF	1	0000 UTC 12 May 2015	IFS Cycle 41r1	2.10	https://www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model
	2	0000 UTC 8 Mar 2016	IFS Cycle 41r2 (O1280)	1.31	Same as above
	3	0000 UTC 22 Nov 2016	IFS Cycle 43r1	2.58	Same as above
	4	0000 UTC 17 Jul 2017	IFS Cycle 43r3	5.22	Same as above
NCEP	1	0000 UTC 14 Jan 2015	TIN14–46 (T1534)	−4.12	http://www.nws.noaa.gov/om/notification/tin14-46gfs_cca.htm
	2	0000 UTC 11 May 2016	TIN16–11 (4DEnVar)	7.37	http://www.nws.noaa.gov/os/notification/tin16-11gfs_gdasaaa.htm
	3	0000 UTC 19 Jul 2017	SCN17–67 (NEMSIO)	0.81	http://www.nws.noaa.gov/os/notification/scn17-67gfsupgrade.htm
UKMO	1	0000 UTC 21 Nov 2016	PS38 (satellite obs)	4.75	https://www.metoffice.gov.uk/research/news/2016/latest-met-office-global-model-improvements
	2	0000 UTC 7 Sep 2017	PS39 (10-km resolution)	2.82	https://www.metoffice.gov.uk/research/news/2017/increased-resolution-of-global-forecast-models

the interval from zero (worst value) to one (best value). An advantage of the minmax approach is that it allows combining any number and type of quantitative PAMs. At first, the focus was on a single global OFS to encapsulate the entire scorecard into one number, but it soon became clear that subsetting the NAMs into a series of SAMs could provide useful diagnostic information. Subsequently, Hoffman et al. (2017a) introduced the empirical cumulative density function (ECDF) transform to map each PAM value to the unit interval. Computationally, this requires ranking the PAMs within the reference sample. Recently, Boukabara et al. (2018) compared minmax to ECDF SAMs within the context of an intercomparison of observing system simulation experiments (OSSEs) and observing system experiments (OSEs) and found similar results using either normalization. In the present study, a minor reformulation of the ECDF approach and a rescaling of the minmax NAMs is shown to lead to SAMs that are very similar, even though the two normalizations are very different. In many results presented below, results for both the ECDF and rescaled-minmax SAMs are plotted together.

The organization of this paper follows the flow of how SAMs are calculated. This calculation involves three transformations: 1) difference the NWP forecasts and analyses, and calculate the PAMs; 2) normalize the PAMs to create NAMs; and 3) combine the NAMs into SAMs. Figure 2 shows the flow of this process from input grids to PAMs to NAMs to SAMs, as well as the other required datasets—the verification grids and the reference samples of PAMs. (This figure is generic, except, as mentioned above, some normalizations, such as the transformation to a z or t statistic or normalization by tabulated values, do not require reference samples.) The first transformation, the calculation of PAMs from model grids, will be described in section 2. Then, section 3 describes the array of PAMs used in our investigation. The second transformation—normalization—is described in section 4 with an emphasis on the choice of reference sample and normalization technique. Examples

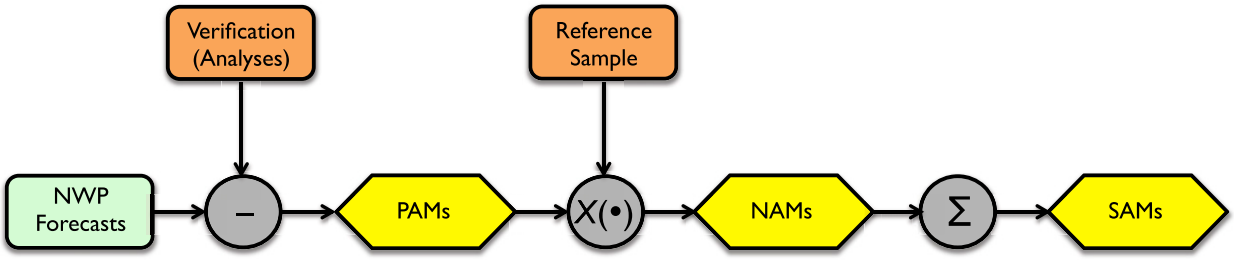

FIG. 2. Flowchart describing the process of transforming forecast grids to PAMs to NAMs to SAMs.

based on the 120-h 500-hPa NHX vector wind RMSE are given in sections 3 and 4. The third transformation involves simple averaging of NAMs over varying subsets to create SAMs. Numerous examples of SAMs are presented and discussed in section 5. Section 6 summarizes and discusses our approach and findings. Technical details are presented in the appendix. In the first section of the appendix, the definitions of the ECDF and minmax normalizations are reviewed. In the second section of the appendix, the rescaled-minmax normalization is described. The third section of the appendix gives a detailed description of how the rank function is used to calculate ECDF NAMs. In the appendix's fourth section, the effective sample size reduction factors, which are used to estimate the SAM uncertainties, are described and calculated.

2. Preliminary data processing

Data used in this study are in verification statistics database (VSDB) format. The VSDB format is used in the NCEP EMC Global NWP model verification package, both for grid-to-grid and grid-to-point comparisons (Zhou et al. 2015; Shafran et al. 2015), and may be used as input to MetViewer (Shafran et al. 2015). Here, we describe those aspects of the VSDB that are related to the grid-to-grid comparisons used in this study.

Operational analysis and forecast grids are exchanged with a number of centers by EMC in GRIB format. If the grids are not already at 2.5° × 2.5° resolution, then they are bilinearly interpolated to this resolution. VSDB files are created by comparing forecasts from each center to the analysis from that center valid at the forecast time. The VSDB files contain the geographic domain means[3] of the quantities needed to compute the desired statistics for different forecast times, vertical levels, domains, variables, and verification times. These quantities are the means of x, x^2, xy, etc., where x and y are fields defined over the domain. The domains used in this study are detailed in Table 2.

The standard WMO verification formulas are conveniently defined by Janousek (2018). Note the following:

- The means are treated as expectations in the statistical formula. Thus, there is no adjustment for degrees of freedom lost in estimating standard deviations and correlations.[4]

TABLE 2. The domains used in this study. For each domain, the first and last longitude (λ_1, λ_L) and latitude (ϕ_1, ϕ_M), the number of longitudes L, the number of latitudes M, and the number of grid points (LM) are listed. Note that 20°N is included in both tropics and NHX domains.

Domain	λ_1	ϕ_1	λ_L	ϕ_M	L	M	LM
Global	2.5°	−90.0°	360.0°	90.0°	144	73	10 512
NHX	2.5°	20.0°	360.0	80.0°	144	25	3600
Tropics	2.5	−20.0°	360.0°	20.0°	144	17	2448
SHX	2.5°	−80.0°	360.0°	−20.0°	144	25	3600

- For grid-to-grid comparisons each grid point is weighted by the cosine of latitude.
- For the vector wind statistics, ordinary multiplications are replaced with dot products in the definitions of the statistical quantities.
- In the calculation of AC, the domain mean anomaly is removed.

EMC exceptions to this standard are that no latitude weighting is used and that the domain mean anomaly is not removed in the AC calculation. These exceptions were made for simplicity before the WMO standard was established.

The VSDB AC scores are calculated using anomalies from the 30-yr (1959–88) climatology of the NCEP–NCAR reanalysis (http://www.emc.ncep.noaa.gov/gmb/STATS_vsdb/). The climatology grids are defined for each day of the year at the four synoptic times: 0000, 0600, 1200, and 1800 UTC. As with the analysis and forecast grids, the original reanalysis 1° × 1° grids are bilinearly interpolated onto 2.5° × 2.5° grids. Then, the grids for each synoptic time and calendar day are averaged over the 30 years. For the upper-air fields used in this study, this climatology is available only at 1000, 700, 500, and 250 hPa for geopotential height and only at 850, 500, and 250 hPa for temperature and wind (components). The 29 February climatology is taken to be the average of the 28 February and 1 March climatology. The AC is calculated as a simple correlation of the forecast and analysis anomalies (i.e., after the climatology for each grid point and variable for the verification day has been subtracted from the forecast and verification). Note that these AC results do not include centering to remove domain mean anomaly errors.

3. Primary assessment metrics

After collecting selected VSDB domain averages and computing the desired statistics, we arrange the results into an array of PAMs with the following dimensions and coordinates listed here in "dimension::coordinate values" format:

[3] These domain means are sometimes referred to as partial sums, but they are really sums divided by the number of grid points in the domain.

[4] It is nontrivial to estimate the degrees of freedom in fields of geophysical variables such as these (Bretherton et al. 1999).

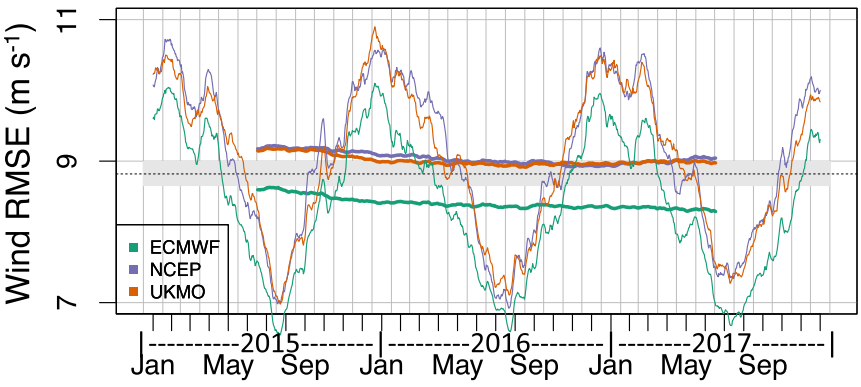

FIG. 3. The evolution of 120-h 500-hPa NHX vector wind RMSE. Over the study period (2015–17) this metric—ECMWF, mint; NCEP, lavender; and UKMO, ochre—has been filtered with a centered 31-day (light lines) and a 365-day (heavy lines) moving average [referred to as MA(31) and MA(365) in the text]. A 95% uncertainty band is plotted for the 365-day filter under the null hypothesis that there is no difference between centers for this metric. Here, $\gamma = 0.6644$ for the 31-day filter and 0.6611 for the 365-day filter.

1) forecast time::24, 48, 72, 96, 120, 144, 168 h;
2) level::250, 500, 700, 850, 1000 hPa;
3) domain::NHX, SHX, tropics (see Table 2);
4) variable::height Z, temperature T, wind V;
5) statistic::AC, RMSE, AME (i.e., the absolute value of bias);
6) verification time::every 24 h at 0000 UTC during 2015–17 for a total of 1096 days; and
7) center::ECMWF, NCEP, UKMO.

Since the PAMs are calculated with respect to the corresponding center's analysis, we do not include the 0-h forecast time. We use AME since it is necessary to take the absolute value of the bias before normalizing it. The available PAMs for both RMSE and AME are mostly complete, including all possible combinations of coordinates. However, some of the possible AC PAMs are not calculated, in keeping with standard NCEP practice. As a result, ACs are missing for 850-hPa height, and for 700- and 1000-hPa temperature and wind. For some purposes it is convenient to reshape the verification time dimension, for example into day, month, and year dimensions.

As an example of a PAM, consider Fig. 3, which displays the evolution of 120-h 500-hPa NHX vector wind RMSE over the study period (2015–17) for the three centers filtered with a monthly (centered 31 day) moving average [MA(31)[5]] and with an annual (centered 365 day) moving average [MA(365)]. In plots of metrics, like this plot, a thin black horizontal line indicates the overall mean value and the gray band about this mean value corresponds to the 95% confidence interval of a Student's t test for the null hypothesis that all the metrics plotted, in this case the MA(365) values, are random draws from the same distribution. Correlations are accounted for by reducing the sample size n to an effective sample size $n^* = n\gamma$, where the reduction factor γ is determined as explained in the appendix, section d. In Fig. 3 the reduction factor for the MA(31) PAMs is $\gamma = 0.6644$, so that there are effectively 20 degrees of freedom in a 30-day sample. Since the two γ values are essentially equal, the MA(31) 95% uncertainty band (not drawn in the figure) is almost 3.5 ($\sqrt{365/31}$) times the width of the MA(365) uncertainty band plotted.

As seen in Fig. 3, typical values for the 120-h 500-hPa NHX vector wind RMSE are roughly 9 m s^{-1}, but there is a very distinctive annual cycle with amplitude of almost ± 2 m s^{-1}. An annual cycle is less pronounced or not apparent at all for other PAMs. In the present example, during the Northern Hemisphere winter, vector winds are stronger than during other seasons and errors are larger. In part, the same displacement error of a synoptic system will lead to larger errors when the field amplitudes are larger, as is the case for the winter NHX vector wind. Relative errors (e.g., errors as a percent) would not be expected to vary this much seasonally. It is clear in Fig. 3 that for this PAM, ECMWF is more skillful than NCEP, which is approximately as skillful as UKMO, with an increase in RMSE of about 0.5 m s^{-1} from ECMWF to NCEP and UKMO. However, improvements in this PAM are barely apparent during the study period and then only after annual averaging. It is questionable whether these improvements rise to the level of statistical significance. On the one hand, the reductions in RMSE for each center are smaller than the width of the uncertainty band. However, there is a clear linear decrease in RMSE for all three centers during

[5] Here, MA(q) indicates the moving average with span (or order) q.

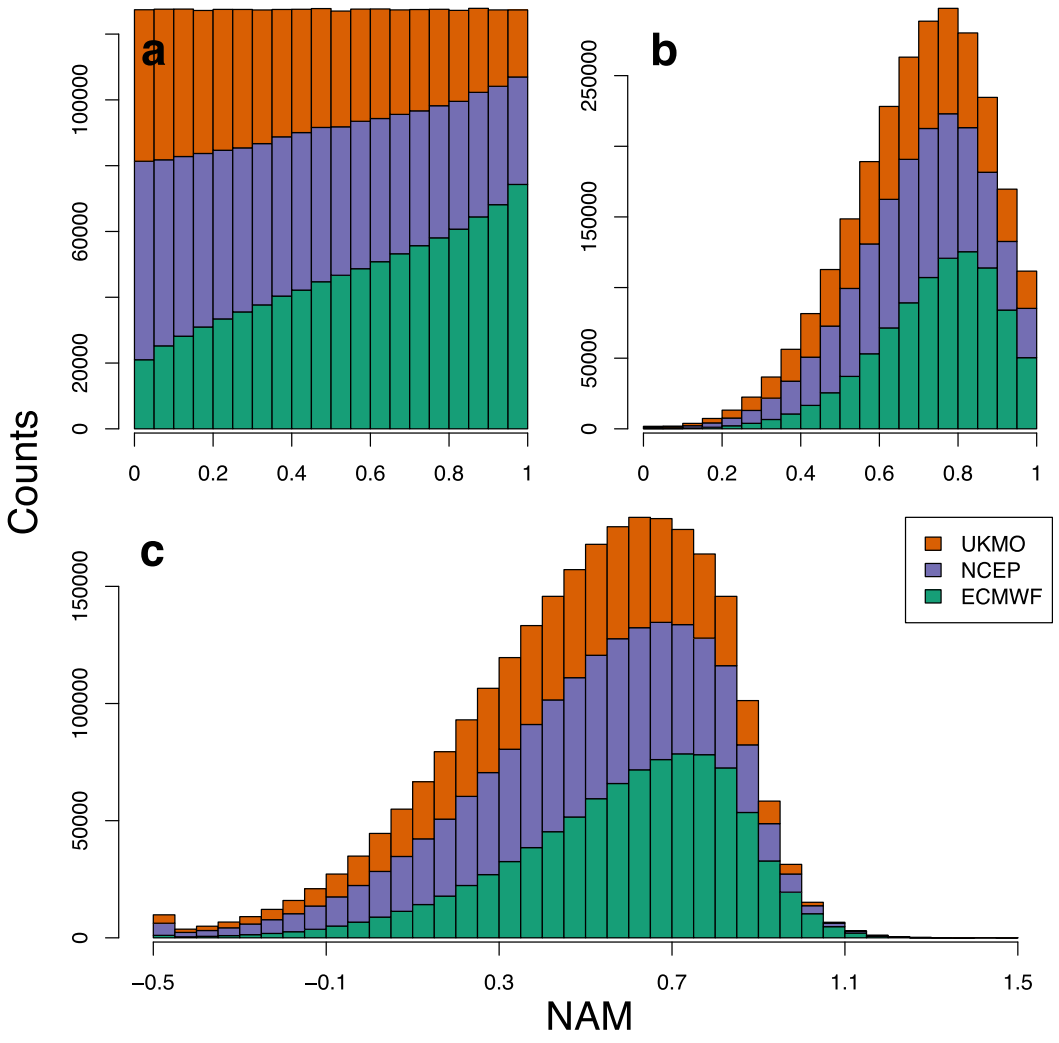

FIG. 4. Histograms of NAMs for the NWP centers for the All reference sample for (a) ECDF, (b) minmax, and (c) rescaled-minmax normalizations. Bin size is 0.05. Slight deviations from a perfectly flat histogram for the ECDF normalization are due to ties.

the first half of the period. This trend continues for ECMWF for the entire period but flattens out for NCEP and UKMO. In sum, Fig. 3 illustrates (i) an annual cycle in skill and (ii) different levels of skill for different centers, as well as (iii) demonstrating only very small improvements in skill during 2015–17. While the first point is true for only some PAMs, the second and third points are true for most PAMs.

4. Normalized assessment metric methodology

SAMs depend on the normalization method, the reference sample used in the normalization, and the subsets of NAMs averaged. In the results discussed here the normalization is either the ECDF normalization or the rescaled-minmax normalization (sections a and b in the appendix). For most of the results one of two reference sample definitions is used. The first—the All reference sample—includes all centers and all verification times during 2015–17. The second—the ByCenter reference sample—includes all verification times for each center. The subsets of NAMs averaged vary in the figures of SAMs.

The details for four normalization methods are described in the appendix. In summary, the ECDF normalization uses the empirical CDF of the reference sample to map each PAM to the probability of randomly choosing a PAM of equal or lower forecast skill.[6]

[6] Since ECDF NAMs are probabilities and ECDF SAMs are averages of probabilities, we may express them in terms of fractions or percentages.

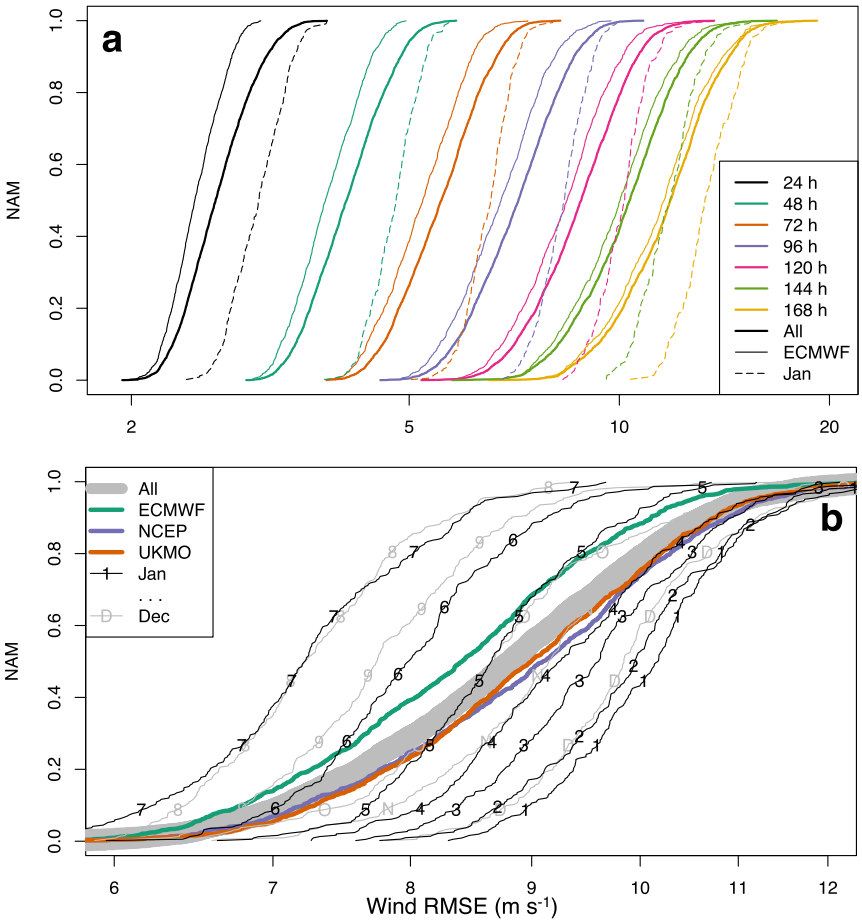

FIG. 5. ECDF transfer functions from PAMs to NAMs for the 500-hPa NHX vector wind RMSE for (a) 24–168-h forecasts and for (b) the 120-h forecast alone for three different reference samples: All, ByCenter, and ByMonth. Different line types, weights, and colors are used to indicate the different ECDFs, as indicated in the legend. See text for more details.

Thus, the best PAM is assigned a NAM value of 1 and the worst a value of 0. The other normalizations apply a linear transformation to the PAMs of the form

$$\mathrm{NAM} = a\mathrm{PAM} + b, \tag{1}$$

where a and b are chosen separately for each reference sample. For the minmax normalization the linear transformation maps the best PAM in the reference sample to 1 and the worst PAM in the reference sample to 0, just as in the ECDF case. For the rescaled-minmax normalizations, the linear transformation creates identically distributed NAMs with a mean of 1/2 and variance of 1/12—the same mean and variance as for the ECDF NAMs. A plain normalization is used in section d of the appendix, which is similar to the rescaled-minmax normalization, except that the plain NAMs have a mean of 0 and a variance of 1.

Note that the ECDF and minmax normalizations are quite different, but the rescaled-minmax normalization includes a linear transformation of the minmax NAMs so that the rescaled-minmax NAMs have the same mean (1/2) and variance (1/12) as the ECDF NAMs. As a result, for these two normalizations, if the NAMs are identically, independently distributed, then, according to the central limit theorem, the SAMs will have a normal distribution with a mean of 1/2 and variance of $1/(12n)$, where n is the number of NAMs (i.e., the size of the subset) that are averaged. Since the NAMs are not uncorrelated, we replace n with an effective number of NAMs $n^* = n\gamma$, where γ is a factor that in an approximate way accounts for the limited degrees of freedom for the dimensions averaged over. (See section d in the appendix.) The calculation of γ follows the approach of Bretherton et al. (1999). The value of γ is given in the figure captions and depends on what dimensions are

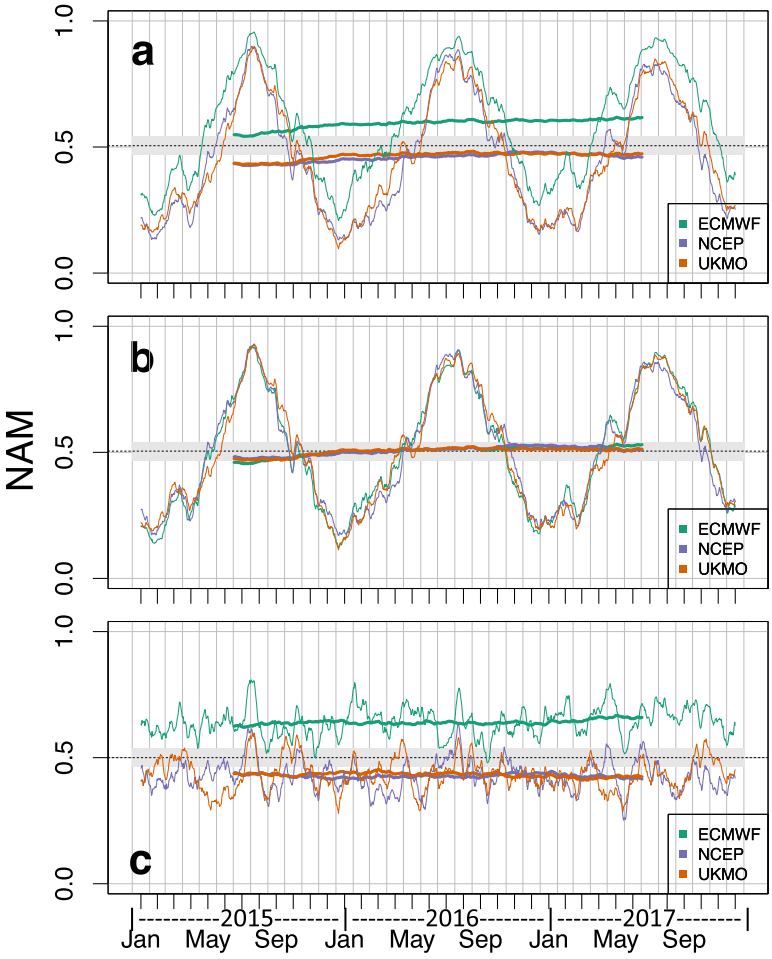

FIG. 6. The evolution of 120-h 500-hPa NHX vector wind RMSE NAMs for three different reference samples: (a) All, (b) ByCenter, and (c) ByMonth. Otherwise, as in Fig. 3.

averaged. By design the ECDF NAMs are identically distributed with a uniform distribution on the unit interval. The distribution of PAMs or linear transformation of PAMs [Eq. (1)] is in general arbitrary, but for cases examined, distributions of RMSE and (1-AC) are well approximated by lognormal distributions and distributions of AME are well approximated by positive halves of normal distributions. Example distributions of NAMs for three normalizations are plotted in Fig. 4.

To explain the concept of the reference sample $\mathcal{R}$, we first consider a single PAM value x, for example, the 120-h 500-hPa geopotential NHX NCEP forecast AC valid 8 August 2016. Here, the PAM type is the 120-h 500-hPa geopotential NHX forecast AC. Reference samples are then all PAMs of this type, for a selection (or all) of the verification times and a selection (or all) of the centers. The corresponding NAM value y will be determined by the relationship of x to $\mathcal{R}$. When using the ByCenter reference sample, there is a different reference sample for each center, and any impacts seen are variations in the skill of that center relative only to itself during the study period. In the present example, we would choose $\mathcal{R}$ to be the set of 120-h 500-hPa geopotential NHX NCEP forecast ACs valid at all verification times. When using the All reference sample there is a single reference sample for all centers, and any impacts seen are variations in the skill of that center relative to itself and all other centers during the study period. In the present example, we would choose $\mathcal{R}$ to be the set of 120-h 500-hPa geopotential NHX forecast ACs valid at all verification times *for all centers*. The All and ByCenter reference samples are used in all of what follows with two exceptions. First, Figs. 5 and 6 include the ByMonth reference sample. For example, the January ByMonth reference sample includes all Januaries and all centers. Second, section d of the appendix makes use of special definitions of the

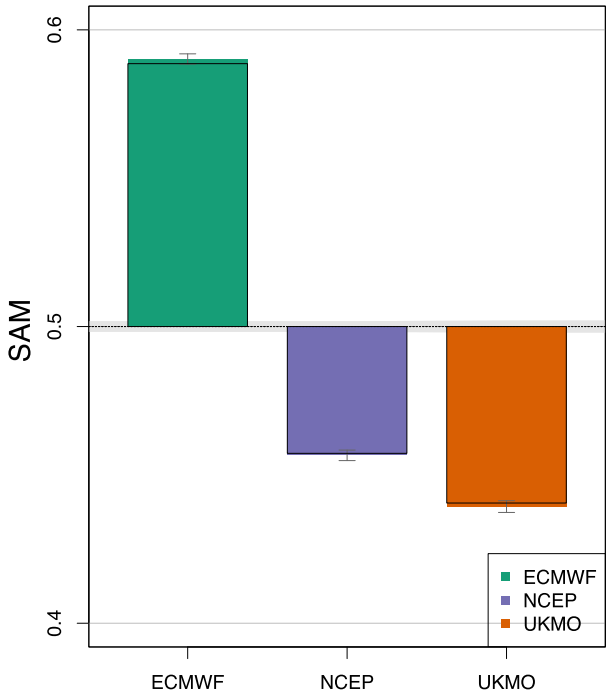

FIG. 7. The global SAM by center during the study period (2015–17). In this and all plots of SAMs, a 95% uncertainty band is plotted around 1/2 under the null hypothesis that there is no difference between centers for this metric. In addition, in this and subsequent plots of this type, the estimated uncertainty at the 95% level is indicated by small error bars at the ends of the color bars, which are anchored at the expected value (1/2), and both the ECDF (colors) and rescaled-minmax normalization (black outline) are shown. In this and immediately following figures, the SAMs are for the All reference sample. Here, $\gamma = 0.1084$.

reference samples that include only a single month and a single center (Monthly), or a single month and all centers (MonthlyAll).

Figure 5 shows examples of ECDF transfer functions (see section c of the appendix for details) from PAMs to NAMs, all for the case of the 500-hPa NHX vector wind RMSE. In Fig. 5a the colors indicate forecasts of 24–168 h. The thick lines for the All reference sample clearly show how the distribution of error increases in magnitude and spread with forecast length. (Note the use of a logarithmic x axis.) As examples, Fig. 5a also shows the ECDF for the ECMWF ByCenter reference sample and for the January ByMonth reference sample. These examples were chosen for clarity since ECMWF errors are the smallest among the centers and January errors are the largest among the months. Fig. 5b focuses on the 120-h forecast and shows the variation with center and month for the ByCenter and ByMonth reference samples. (The magenta lines from Fig. 5a are repeated in Fig. 5b but with different line styles.) In Fig. 5b, the ordering of the pdf distributions by center (the ByCenter ECDFs) shows that for this PAM ECMWF $<$ NCEP $\sim$ UKMO (and the reverse sense of the inequality holds for forecast skill). Similarly, the ordering of the ByMonth ECDFs shows the annual cycle of wind speed in the NHX is reflected in the wind errors, with January having the largest errors and July and August having the smallest errors.

Figure 6 illustrates how different reference samples affect the resulting NAMs by converting the 120-h 500-hPa NHX vector wind RMSE PAMs of Fig. 3 into NAMs using ECDF transfer functions for the three different reference samples depicted in Fig. 5. The patterns for All (Fig. 6a) and ByCenter (Fig. 6b) are very similar, but in Fig. 6a, where ECMWF is compared to all centers, the NAMs for ECMWF are larger since ECMWF is more skillful and the range of the ECMWF monthly averaged NAMs is compressed upward since the other centers contribute many lower skilled forecasts to this reference sample. In both Figs. 6a and 6b it is difficult to discern a trend in skill for the monthly average NAMs. In terms of annual averages, there are slight improving trends, more so during the first half of the period, in both Figs. 6a and 6b. In Fig. 6c the seasonal signal is removed by comparing PAMs only to other PAMs of the same calendar month. Note that removing the annual cycle in this way emphasizes the ECMWF forecast superiority for this PAM. The concordance of the three monthly curves in Fig. 6b is striking. The variations on weekly to seasonal time scales in Fig. 6b match up quite well from center to center. Thus, normalizing each center separately results in nearly identical behavior in response to changing atmospheric conditions.

5. Summary assessment metric results

Figure 7 shows the global SAM by center during the study period. Here, all available NAMs for each center have been summed and the All reference sample is used to compare the centers. The global SAMs indicate that ECMWF $>$ NCEP $\gtrapprox$ UKMO in terms of forecast skill. Under the null hypothesis that there is no difference between centers, all the SAM values would be 1/2. Therefore, we refer to the difference relative to 1/2 as the SAM impact. A value of 3/4 would indicate a strong positive impact (improvement in forecast skill relative to the reference sample), and a value of 1/4 would indicate a strong negative impact. Given the very large number of NAMs that contribute to these SAMs, the uncertainty of the estimates (error bars at the ends of the color bars and gray horizontal band around 1/2) are very small even though we have accounted for correlations in estimating the effective sample size. In Fig. 7 the

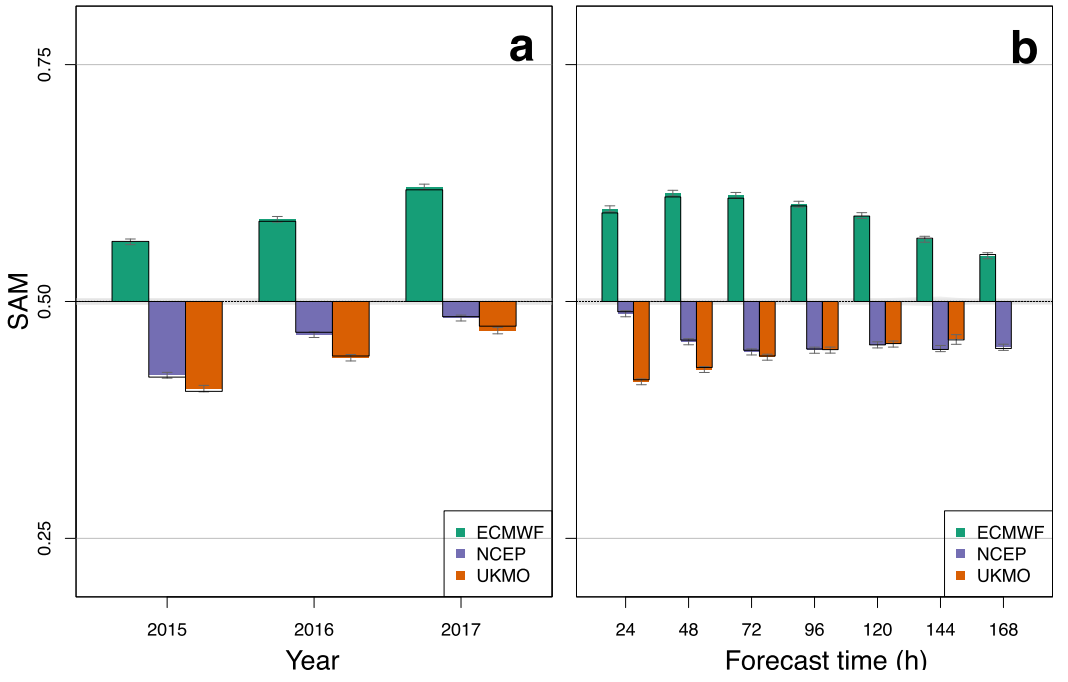

FIG. 8. Variation of SAM by center and by (a) year and (b) forecast time. Otherwise, as in Fig. 7. [Reference sample is All; $\gamma = 0.1085$ in (a), and $\gamma = 0.2290$ in (b).]

reduction factor for the number of degrees of freedom is $\gamma = 0.1084$. (Please refer to section d of the appendix.) Note that Fig. 7 shows results both for the ECDF normalization (color bars) and the rescaled-minmax normalization (black outlines). Here and in other figures the differences between the results using these two normalizations are in some cases statistically significant (i.e., larger than the uncertainties) but are in most cases very small, for example, compared to differences due to other factors.

Figure 8a plots the variation of SAM by year for each center. All three centers are improving and at approximately the same rate. Figure 8b plots the variation of SAM with respect to forecast time by center. Overall, SAM impacts decrease with forecast time. This is consistent with the expectation that at long enough forecast times all SAMs will be close to ½ since each model forecast should be a random draw from the climatological distribution. Comparing NCEP and UKMO in this plot, we see that initially NCEP skill relative to the other centers decreases with forecast time, while UKMO skill increases, and the two are roughly equivalent beyond 48 h. Note that in Fig. 8b there are differences in error evolution with forecast time due to both differences in initial state errors and due to differences in NWP model errors. Similar comparisons in our OSE and OSSE studies (Boukabara et al. 2016, 2018; Hoffman et al. 2017a) have a more distinctive look since, as a single model is used, the NWP model errors are the "same." In such comparisons SAM impacts quickly decay with forecast time. That is, there are large differences in SAMs between impact experiments at the initial time and much smaller differences at long forecast times. This is seen in Fig. 8b for UKMO and beyond 48 h for ECMWF, but not for NCEP, suggesting that model errors or inconsistencies between the analysis and model are dominating the NCEP results.

Figure 9 is similar to Fig. 8, but shows how SAMs vary for each center as a function of level, domain, variable, and statistic. By level (Fig. 9a), there are greater differences between centers at lower levels. By domain (Fig. 9b), NCEP, and to a lesser degree UKMO, are relatively good in the tropics. By variable (Fig. 9c), UKMO is good at geopotential height but not at temperature. Conversely, NCEP is good at temperature but not at geopotential height. By statistic (Fig. 9d), note that ECMWF does not perform as well for AME in a relative sense as it does for AC and RMSE.

Similar to Fig. 8b, we further partition the NAMs into SAMs that depend on domain (Fig. 10a) or year (Fig. 10b). Typically, plots of this sort do not show interactions; that is, we expect Fig. 10a to be a product of Figs. 8b and 9b. This is true for NCEP: the impacts for the domains in Fig. 10a are similar in shape to the impacts in Fig. 8b, but the magnitude of these impacts varies with the impact by domain seen in Fig. 9b.

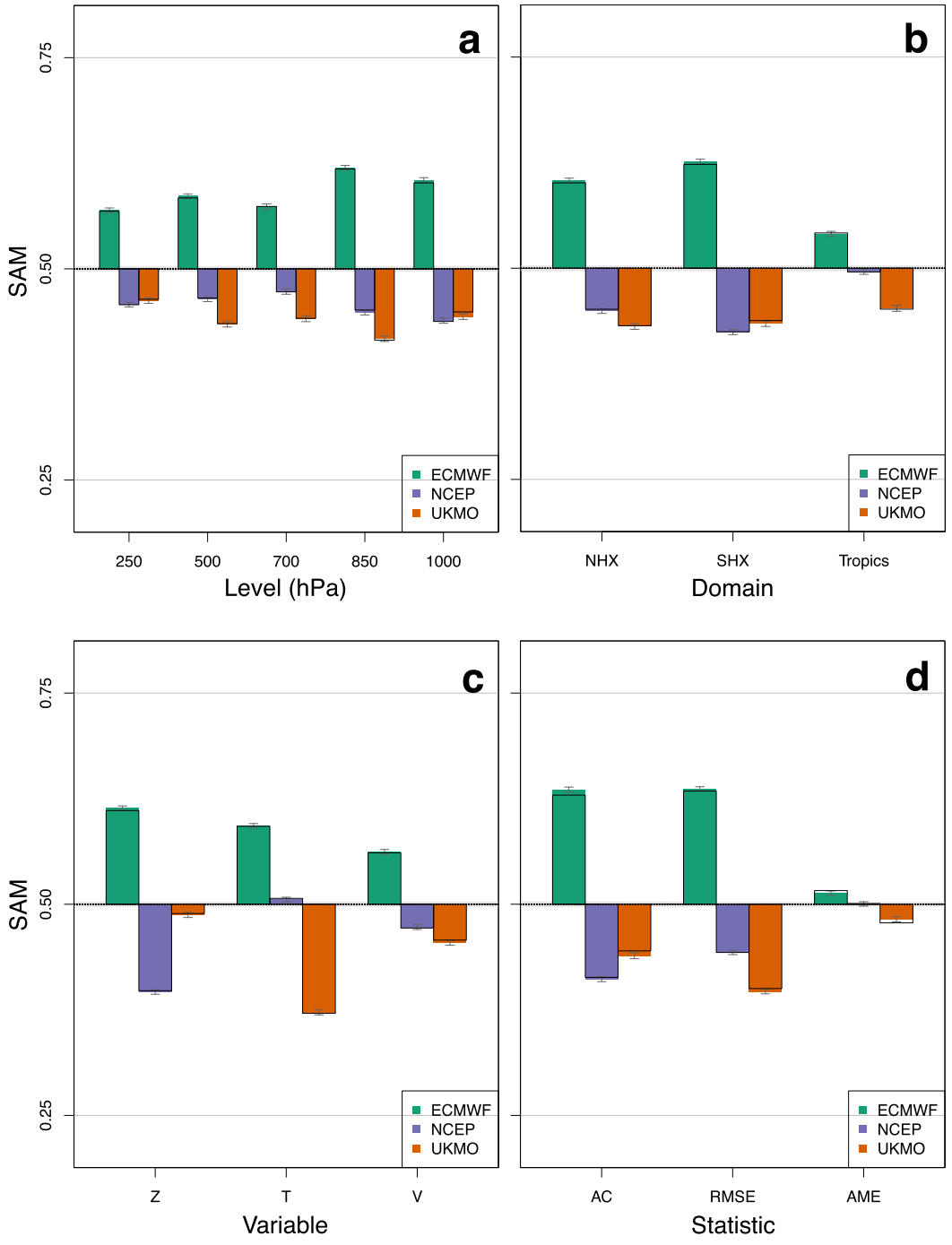

FIG. 9. Variation of SAM by center and by (a) pressure level, (b) geographic domain, (c) variable, and (d) statistic. Otherwise, as in Fig. 7. [Reference sample is All; $\gamma = 0.1972$ in (a), $\gamma = 0.1090$ in (b), $\gamma = 0.1348$ in (c), and $\gamma = 0.1375$ in (d).]

For ECMWF and NCEP, the impacts in the tropics have different evolutions with the result that the SAM impacts for the tropics for ECMWF and NCEP cross close to 48 h. In Fig. 10b, all three centers improve year by year. In a relative sense, NCEP in 2016 is much better than in 2015. The improvements in NCEP performance in the tropics and after 2015 are presumably due to the May 2016 4DEnVar implementation (upgrade NCEP-2 in Table 1).

To eliminate the effect of one center on another, we now turn to SAMs based on the ByCenter reference

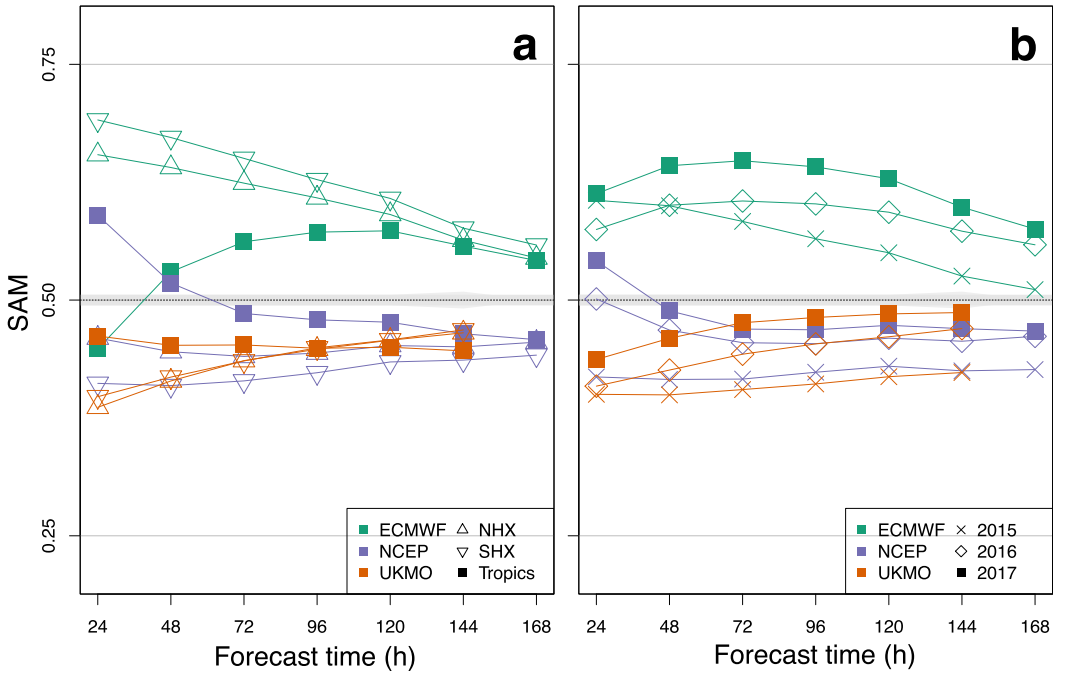

FIG. 10. Variation of SAM by center and forecast time and by (a) geographic domain and (b) year. Otherwise, as in Fig. 7. [Reference sample is All; $\gamma = 0.2302$ in (a), and $\gamma = 0.2292$ in (b).]

samples. Even though we plot all centers together, each center is only compared to itself in the next several figures. Since the SAM impacts are smaller using the ByCenter reference samples, the *y* axis has been changed for these plots. Figure 11a shows the variation of SAM by center from 2015 to 2017 and is analogous to Fig. 8a. Again, it is clear that all three centers are improving regularly from year to year. Figure 11b reiterates what

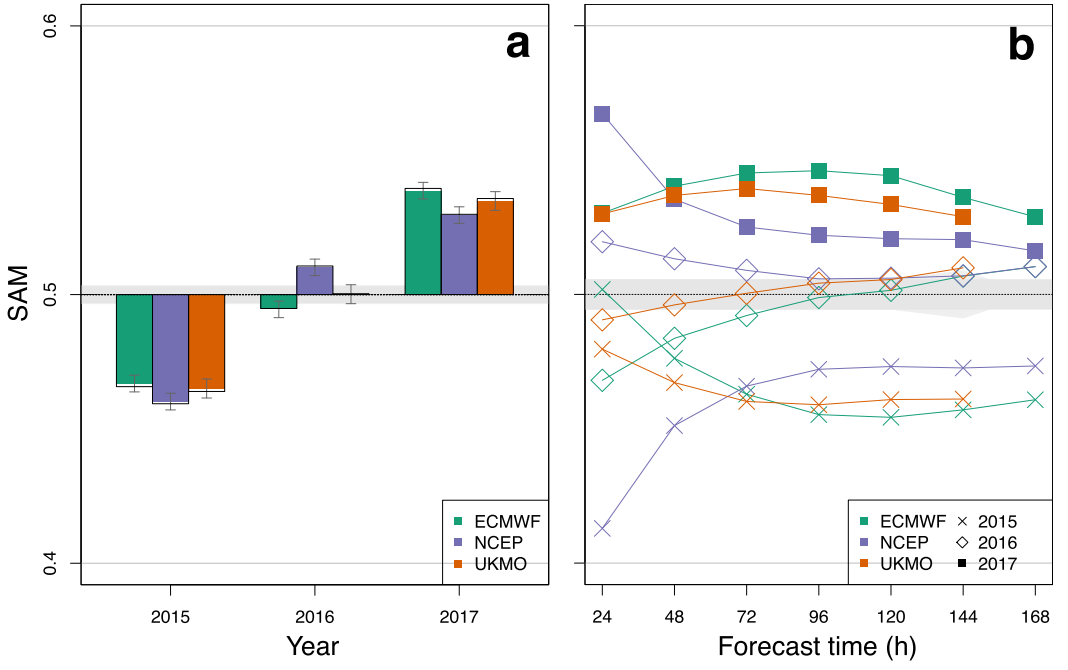

FIG. 11. The variation of SAM (a) by center from 2015 to 2017 and (b) with forecast time for each year from 2015 to 2017. In this and immediately following figures, the SAMs are for the ByCenter reference samples. Otherwise, as in Fig. 7. [Here, $\gamma = 0.1085$ in (a) and $\gamma = 0.2292$ in (b).]

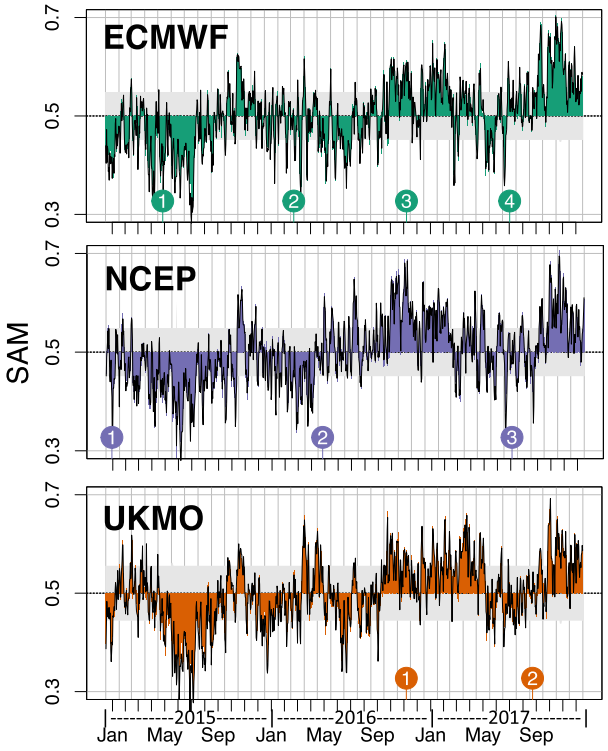

FIG. 12. The evolution of day-to-day SAM for each of the three centers. The major upgrades, marked by the circled numbers along the bottom of each panel, are detailed in Table 1. In a manner similar to the other figures, the ECDF SAMs are plotted in color fill and the rescaled-minmax ECDFs are plotted as a black line. (Reference sample is ByCenter; $\gamma = 0.1642$.)

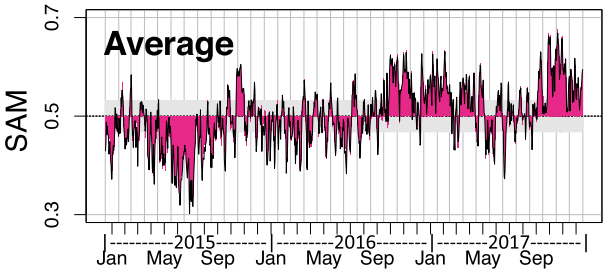

FIG. 13. The evolution of day-to-day SAM averaged over the three centers. Otherwise, as in Fig. 12. (Reference sample is ByCenter; $\gamma = 0.1299$.)

was demonstrated by Fig. 10b (i.e., that the NCEP increase of skill is mostly due to short forecast times).

Figure 12 displays the day-to-day SAMs for each of the three centers. All cases have large day-to-day variability as well as weekly and 30–90-day variability. This variability partially masks the year-to-year trends seen in Fig. 11a. The 30–90-day variations in SAM impact look similar for each center. However, when the day-to-day variability is added, the average of all the centers (Fig. 13) shows a somewhat reduced-amplitude version of the 30–90-day variations seen in Fig. 12. The major upgrades for each center, marked by the circled numbers along the bottom of each panel do not seem to be directly associated with increases in forecast skill as seen in the SAM impacts, except for the following two cases. First, considering the time series of SAM before and after upgrade NCEP-2, it seems clear that the upgrade to the 4DEnVar was associated with a step up in forecast skill. The changes in SAM ($\times 100$; i.e., as probabilities in percent) due to the upgrades are listed as Δ in Table 1. For upgrade NCEP-2 (TIN16–11; 4DEnVar), the "SAM impact" is $\Delta = 7.37\%$. In other words, the effect of implementing the 4DEnVar is a 7.37% increase in the probability of improved skill relative to a randomly chosen forecast metric from 2015 to 2017. This is the largest SAM impact during the study period. Second, for the UKMO-1 upgrade (PS38; $\Delta = 4.75\%$), numerous improvements to how satellite observations are used resulted in a fairly clear increase in SAM. Next, note the two upgrades associated with higher resolution. Upgrade NCEP-1 (TIN14–46; $\Delta = -4.12\%$) from T574 to T1534 spectral resolution has an unexpected negative Δ. This is probably due to sampling effects—the NCEP-1 implementation date is only 14 days from the start of the study period—and would likely be reversed if our sample included 2014. Upgrade UKMO-2 (PS39; $\Delta = 2.82\%$) from 17- to 10-km resolution does appear to be associated with some improvement in skill. For ECMWF there are upgrades every eight or so months, with the latest one to IFS Cycle 43r3 associated with the greatest improvement (upgrade ECMWF-4; $\Delta = 5.22\%$). While there is a definite improvement with time for ECMWF, it is hard to see the connection to the actual upgrades.

Note that for upgrade UKMO-1 the improvements seem to precede the date of the upgrade. This could be caused by NCEP receiving the upgraded model results during a parallel testing period before the official upgrade. Or, more likely, as this upgrade became operational during the month of November, this is due to the yearly improvement of forecast scores during the Northern Hemisphere fall season.

To see the similarities and differences in the 30–90-day variability in forecast skill more clearly, Figs. 14 and 15 average the results of Fig. 12 in the time dimension to smooth out the day-to-day variability. Figure 14 plots the MA(31) and MA(365) SAM results for each of the three centers. To the eye, the ECMWF, NCEP, and UKMO MA(31) SAMs are all tracking each other quite well, except during the first half of 2016. This is consistent with the statement that the atmospheric state is a major determinate of variations of forecast skill. The improvements in MA(365) SAM in Fig. 14 show

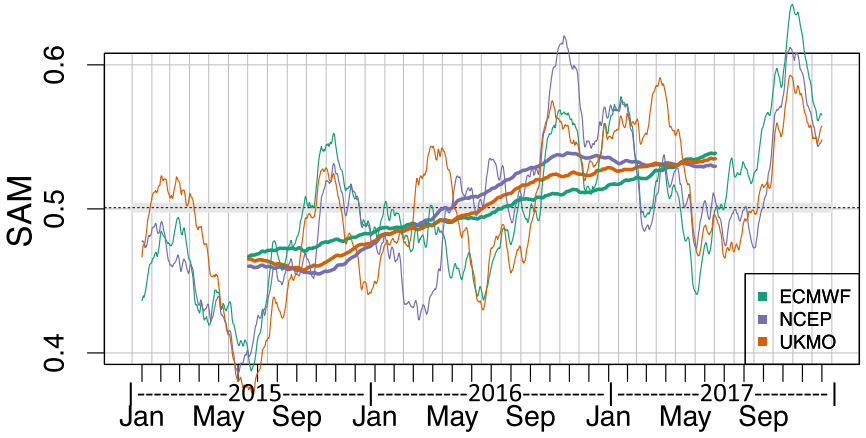

FIG. 14. The evolution of day-to-day SAM. Otherwise, as in Fig. 3. (Reference sample is ByCenter; $\gamma = 0.1091$ for the 31-day filter and and 0.1085 for the 365-day filter.)

improvements that are very statistically significant, with increases in SAM an order of magnitude larger than the 95% uncertainty band. Figure 15 plots SAM by calendar month averaged over the 3 years of the study period for each center. The annual cycle is similar for each center, especially during the July–November improvement in forecast skill.

6. Concluding remarks

Often skill of a global NWP system is tracked in terms of a few key primary assessment metrics such as the 500-hPa geopotential anomaly correlation or the 250-hPa wind RMSE. A significant challenge in this approach is that focusing on an individual PAM (i.e., a particular statistic for a given forecast time, level, domain, and variable) may ignore other important aspects of forecast skill. This is one reason to consider summary assessment metrics when verifying (or validating and tuning) NWP forecast and data assimilation systems. As well as avoiding the problems of focusing on just a few PAMs, the use of SAMs increases statistical significance and enables one to explore various aspects of forecast skill. Here, SAMs are averages of normalized assessment metrics, and each NAM corresponds to a single PAM. For example, the PAM for the 120-h 500-hPa geopotential NHX NCEP forecast AC valid 8 August 2016 is converted to the corresponding NAM using a normalization that is based on a reference sample of the 120-h 500-hPa geopotential NHX forecast AC PAMs for all valid verification times. While a first-order verification might consider "global" SAMs that combine all NAMs for each center, subsequent validation might consider SAMs for various subsets, typically along different dimensions (e.g., for each year, each forecast time, each month, or each calendar date). For example, Fig. 12 uses SAMs to track the progress in NWP centers' skill day by day during 2015–17. Each center is seen to improve its SAM over the three years 2015–17. The connection between specific upgrades to global NWP systems and improvements in forecast metrics are now difficult to detect because the improvements are small compared to variations in atmospheric predictability. Also, after a major upgrade, there can be a period of minor fixes that improve performance, so we should not

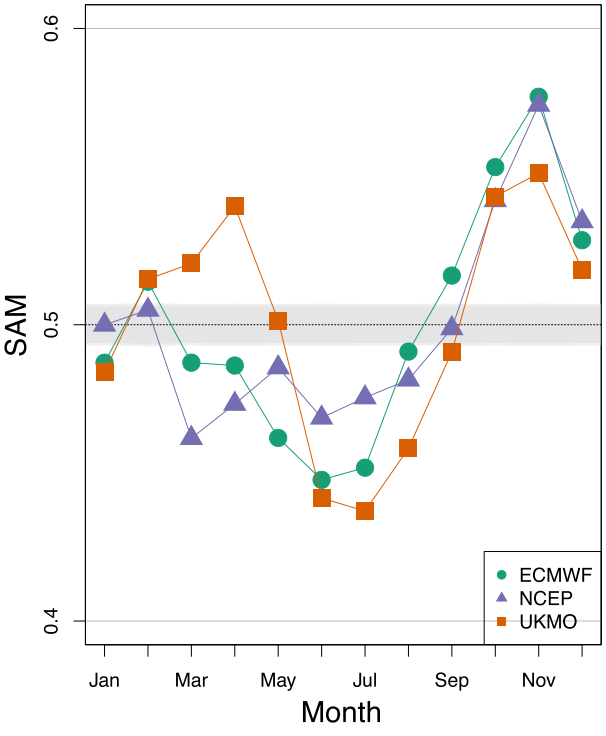

FIG. 15. The variation of SAM by center and by calendar month averaged over the study period (2015–17). (Reference sample is ByCenter; $\gamma = 0.1090$.)

anticipate a step function response to published upgrades. Nevertheless, the use of SAMs improves the signal-to-noise ratio and clear improvements in SAM are related to the ECMWF July 2017 upgrade to IFS Cycle 43r3, the NCEP May 2016 replacement of the 3DEnVar system with the 4DEnVar system, and the UKMO November 2016 (PS38) introduction of improved use of satellite observations.

In general, global NWP forecast skill for the PAMs included in this study is improving at a rate much slower than previously, and long periods are necessary to demonstrate impacts even when using the SAM approach. In future work, it might be interesting to include other centers and to add PAMs for relative humidity and precipitation, forecast variables for which there is currently major room for improvement.

Acknowledgments. The authors thank the many colleagues and collaborating scientists who contributed by their interactions, peer reviews, and suggestions. We gratefully acknowledge support for this work provided by NOAA, including support provided under the auspices of the cooperative institutes through Cooperative Agreements NA14NES4320003 for CICS and NA15OAR4320064 for CIMAS. A preliminary version of this work was presented by Hoffman et al. (2017b).

APPENDIX

Methods

a. Summary assessment metric method

In prior research, we have calculated SAMs using both ECDF and minmax normalizations. Originally, in what we termed ''overall'' scores, a minmax normalization was used (e.g., Boukabara et al. 2016). Later, we proposed the alternative use of an ECDF normalization and applied it to the OSEs of Boukabara et al. (2016) (Hoffman et al. 2017a) and to the 2015 skill scores from several global NWP centers (Hoffman et al. 2017b). Now, we use both the ECDF normalization and a rescaled-minmax normalization. These two normalizations are comparable and we find that they produce very similar SAM results even though the normalizations themselves are quite different. The reason for this is that for either the ECDF or the rescaled-minmax normalizations, the NAM expected value and variance are 1/2 and 1/12. Then, assuming the NAMs are uncorrelated, as a consequence of the central limit theorem, the SAMs should have a normal distribution with mean 1/2 and variance $1/(12n)$ under the null hypothesis (H0). Here, n is the number of NAMs (size of the subset) that are averaged. Since the NAMs are not uncorrelated, n must be replaced by an effective value as discussed below in section d of this appendix.

Each SAM depends on the normalization method, the reference sample used in the normalization, and the subset Ω of NAMs averaged. Usually, we consider a variety of subsets and hold the normalization and reference sample fixed. In this case, the transformation from NAMs to SAMs (the summation symbol in Fig. 2) is given by

$$\mathrm{SAM}_{\Omega} = n^{-1} \sum_{i \in \Omega} \mathrm{NAM}_i, \tag{A1}$$

where n is the number of NAMs in the subset. In all cases the normalization is specific to each PAM type, that is, to each individual forecast time, level, domain, variable, and statistic (e.g., all the 120-h 500-hPa geopotential NHX forecast ACs). For each PAM type, the reference sample $\mathcal{R}$ includes the PAMs for all verification times either for all centers (the All reference sample) or for each center (the ByCenter reference sample).

For a PAM like AC, where increasing values are better, the ECDF normalization is given by

$$\mathrm{NAM} = \frac{\mathrm{rank}(\mathrm{PAM\ in}\ \mathcal{R}) - 1/2}{\mathrm{size}(\mathcal{R})}, \tag{A2}$$

and the minmax normalization is given by

$$\mathrm{NAM} = \frac{\mathrm{PAM} - \min(\mathcal{R})}{\max(\mathcal{R}) - \min(\mathcal{R})}. \tag{A3}$$

The minmax NAMs are in the range [0, 1], with 0 being worst and 1 best. This also holds for ECDF NAMs, but with a slightly reduced range as explained below in section c of this appendix. For a PAM-like RMSE, where increasing values are worse, Eqs. (A2) and (A3) are applied to the negative of the PAM values. Details of the ECDF formulation are given below.

b. Rescaling the minmax NAMs

Within the context of SAMs, the *impact* is the difference between the calculated SAM and its expected value under the null hypothesis that there is no effect due to subset. Thus, for n sufficiently large, which is the case in all examples presented in this paper, ECDF SAM values of 0.75 and 0.25 would represent very large positive and negative impacts, respectively, since under H0 the expected value is 0.5 and the variance scales with $1/n$. For minmax SAMs the expected value (and variance) under H0 varies with subset. For example, the 2015 mean values of minmax SAMs for AC and RMSE are 0.70 and 0.63. To avoid the resulting difficulties in

interpretation of the minmax SAMs, we rescale the minmax NAMs. The first goal of the rescaling is to make all the NAMs similar in terms of having a shared mean and variance. For this purpose the rescaled-minmax NAMs, denoted NAM′, are defined by

$$\mathrm{NAM}' = \left(\frac{\mathrm{NAM} - \mu_S}{\sigma_S}\right)\sigma_T + \mu_T, \qquad \text{(A4)}$$

where μ_S and σ_S are the estimated mean and standard deviation, respectively, of the NAMs in the reference sample, and μ_T and σ_T are the target mean and standard deviation, respectively, of the NAM′ reference sample. A second goal is to make the rescaled-minmax SAMs directly comparable to the ECDF SAMs. Therefore, $\mu_T = 1/2$ and $\sigma_T^2 = 1/12$, which are the expected values for ECDF NAMs. There are a number of advantages to rescaling the minmax NAMs and one disadvantage. The advantages include the following.

- The ECDF and rescaled-minmax SAM impacts have the same definition, and the baseline for the impact is always 1/2.
- The ECDF and rescaled-minmax SAMs are very similar and provide a useful indication that the sensitivity of SAM results to the normalization method is fairly weak.
- The ECDF and rescaled-minmax SAMs are both averages of random variables with the identical mean and variance. Under the null hypothesis H0, both the ECDF and rescaled-minmax SAMs have an expected value of 1/2 and an expected variance of $1/(12n^*)$, where n^* is the effective number of degrees of freedom in the NAMs that are averaged. (See section d of this appendix.) As n^* increases, the distribution of SAMs will approach a normal distribution.

The disadvantage is that the rescaled-minmax NAMs are not restricted to the interval [0, 1] but are transformed according to Eq. (A4) (Fig. 4). However the rescaled-minmax SAMs are very similar to the ECDF SAMs and are generally in the interval [1/4, 3/4].

c. *Definition of NAMs from ECDF*

The ECDF transfer function mapping PAMs to NAMs is defined by a reference sample of n PAM values x_i, $i = 1, \ldots, n$. For the purpose of this discussion, assume that $\mathbf{x}$, the vector of the x_i, is sorted in ascending order. The ECDF is a step function with steps of $1/n$ at the x_i locations. If there are k equal values in the reference sample, then the step at this value is k/n. Figure A1 illustrates this. In Fig. A1, NAM times n (denoted s) is plotted versus PAM (denoted x). Except for the reference sample, the value of s for any x is given by the number of x_i smaller than x. In particular, $s = 0$ (n) and NAM = 0 (1) for x smaller (larger) than all the x_i. At a reference sample value, x_{j+1}, where here, to allow for ties, the subscript $j + 1$ indicates that there are j smaller values in $\mathcal{R}$, the value of s is defined as the midpoint of the riser, which is equal to $j + k/2$, that is, the number of $x_i < x_{j+1}$ plus half the number of ties (including the case of no ties where $k = 1$).

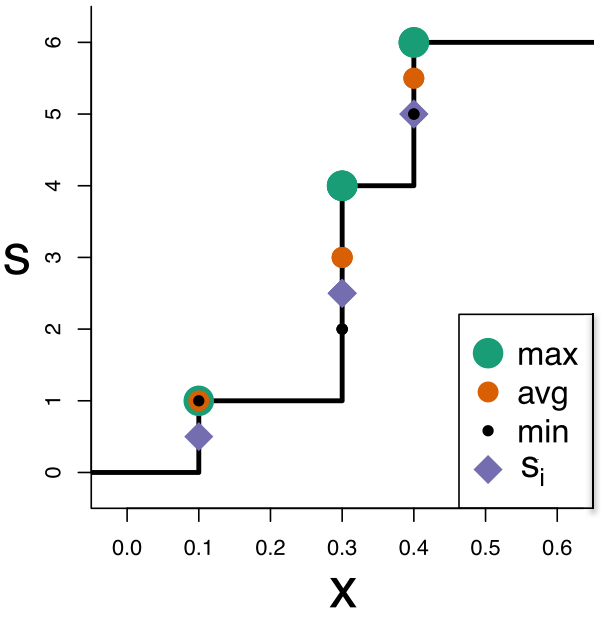

FIG. A1. ECDF example plotting s, the NAM value times n, vs x, the PAM value, where n is the sample size. For this example, the reference sample x_i is (1, 3, 3, 3, 4, 4)/10 and $n = 6$. For any x, s is the number of x_i smaller than x, except when x equals one of the x_i. In this latter case, s is the midpoint of the riser. Three methods of assigning ranks in the case of ties are shown.

An efficient calculation of $\mathbf{s}$, the vector of s values (i.e., the vector of NAMs times n), makes use of the ranking or ordering of $\mathbf{x}$. There are several possible methods of assigning ranks for ties in $\mathbf{x}$. The three methods shown in Fig. A1 are the minimum rank $\check{r}$, the maximum rank $\hat{r}$, and the average rank $\bar{r}$ methods. First, consider assigning ranks to the reference sample $\mathcal{R}$. The value of $\bar{r}(x_{j+1})$ is the average of $i = j + 1, j + 2, \ldots, j + k$ or $j + (1 + k)/2$. Therefore,

$$\mathbf{s}(\mathcal{R}) = \bar{r}(\mathcal{R}) - 1/2. \qquad \text{(A5)}$$

When using $\mathcal{R}$ to calculate $\mathbf{s}$ for an experiment sample $\mathcal{E}$ different from $\mathcal{R}$, start by assigning the values of $\mathbf{s}(\mathcal{R})$ to any values in $\mathcal{E}$ that match a value in $\mathcal{R}$. For all other values in $\mathcal{E}$,

$$\mathbf{s}(\mathcal{E}) = \check{r}(\mathcal{E} + \mathcal{R}) - \check{r}(\mathcal{E}). \qquad \text{(A6)}$$

Consider x, a value in $\mathcal{E}$ that is not in $\mathcal{R}$. Suppose that $\check{r}(x) = 5$ within $\mathcal{E}$ and 15 within $\mathcal{E} + \mathcal{R}$. Then, since x is

TABLE A1. The numbers of degrees of freedom estimated as described in the text for the four normalization methods defined by the equations referenced in the column labeled NAM. In this table (and in subsequent tables), the last nine columns are the dimensions of the PAM array used in this study. The first table row gives d, the sizes of the dimensions. In the dimension column headings: t_f is the forecast length; day, month, and year are the calendar date dimensions of the verification time; and the abbreviations stat., dom., and var. are used for statistic, domain, and variable.

	NAM	Stat.	t_f	Dom.	Var.	Level	Day	Month	Year	Center
d	—	3.00	7.00	3.00	3.00	5.00	31.00	12.00	3.00	5.00
ECDF	A2	2.37	3.31	2.98	2.41	2.75	20.60	11.94	3.00	2.37
Minmax	A3	2.41	3.45	2.99	2.45	2.87	20.83	11.94	3.00	2.42
Rescaled	A4	2.37	3.31	2.98	2.41	2.75	20.62	11.94	3.00	2.37
Plain	A9	2.37	3.31	2.98	2.41	2.75	20.60	11.94	3.00	2.37

larger than 4 elements in $\boldsymbol{\mathcal{E}}$ and 14 in $\boldsymbol{\mathcal{E}} + \boldsymbol{\mathcal{R}}$, we know that x is larger than 10 elements in $\boldsymbol{\mathcal{R}}$. Note that Eq. (A5) uses the average rank while Eq. (A6) uses the minimum rank.

In practice, Eq. (A6) can be applied to all of $\boldsymbol{\mathcal{E}}$, giving a preliminary value of $\mathbf{s}$, say $\mathbf{j}$. Then, if $\boldsymbol{\mathcal{R}}$ is in ascending order, a particular element of $\boldsymbol{\mathcal{E}}$ is in the interval $(x_j, x_{j+1}]$. Therefore, ties can be identified by testing if $\boldsymbol{\mathcal{E}}$ equals $\boldsymbol{\mathcal{R}}(\mathbf{j}+1)$. Finally, for each tie, that element of $\mathbf{j}$ should be replaced by the corresponding value of $\mathbf{s}(\boldsymbol{\mathcal{R}})$.

Note that Eq. (A2) has been modified relative to the definition given by Hoffman et al. (2017a) in two ways to guarantee that the average of the NAMs over the reference sample is exactly 1/2. First, in the numerator 1/2 instead of 1 is subtracted from the rank. Now, in the case of no ties, the mean of the reference sample NAMs must be ½ since the reference sample NAMs are symmetric about 1/2 with a range of $[1/(2n), 1 - 1/(2n)]$. Second, in the case of ties the rank is now the average, not the minimum. Now, ties do not affect the mean of the reference sample NAMs.

d. *Effect of correlations on effective sample size for estimating uncertainties*

Uncertainty intervals for SAMs can be estimated based on (i) the known statistics of the NAMs, (ii) the number n of NAMs averaged, and (iii) the fact that for large n, the SAM will have a normal distribution with a mean and variance related to the mean and variance of the NAMs. For the ECDF and rescaled-minmax normalizations the NAM mean is 1/2, the NAM variance is 1/12, the SAM mean is 1/2, and the SAM variance is $1/(12n^*)$. Here, n^* is the effective sample size to account for the correlations in the sample. We estimate n^* as

$$n^* = n\prod_{i\in\mathcal{D}}\gamma_i = n\gamma_{\mathcal{D}}, \tag{A7}$$

where $\mathcal{D}$ is the set of dimensions averaged, and γ_i is the reduction factor for dimension i. In the main text of the article, an unadorned γ is written for $\gamma_{\mathcal{D}}$, the total reduction factor for $\mathcal{D}$. Each γ_i is given by d_i/ν_i, where d_i is the size of dimension i and ν_i is the number of degrees of freedom in that dimension estimated following Bretherton et al. (1999) using the "eigenvalue formula" [their Eq. (5)], which may be written as

$$\nu_i = d_i^2 \Big/ \sum_{j,k}(C_{jk}^2). \tag{A8}$$

Here, C_{jk} is the jkth element of the correlation matrix $\mathbf{C}$ for dimension i and the denominator of Eq. (A8) is the square of the Frobenius norm of $\mathbf{C}$.

In the present case $\mathbf{C}$ must be estimated after the PAMs are detrended in time, have had the effect of the NWP center removed, and are normalized. This is required to avoid the real signals from inducing strong correlations that would artificially reduce the number of degrees of freedom. For this purpose we calculate NAMs using a special reference sample definition that includes only a single month and a single center (Monthly). The calculated ν for this reference sample agree quite well for the four normalization methods that were applied, as seen in Table A1. The normalizations are those described above plus a "plain" normalization that applies the following simplification of Eq. (A4):

$$\mathrm{NAM} = \frac{\mathrm{PAM} - \mu_S}{\sigma_S}, \tag{A9}$$

where μ_S and σ_S are now the estimated mean and standard deviation of the PAM reference sample. In all three cases the combined reduction factor for all dimensions is approximately 0.09 and for all verification time dimensions about 0.66. For all figures we use the γ_i

TABLE A2. The values of γ_i (%) derived from the plain normalization method.

	Stat.	t_f	Dom.	Var.	Level	Day	Month	Year	Center
γ	78.8	47.3	99.5	80.4	55.0	66.4	99.5	99.9	79.1

TABLE A3. The values of γ_i (%) derived from the ECDF normalization method for different reference samples $\mathcal{R}$.

$\mathcal{R}$	Stat.	t_f	Dom.	Var.	Level	Day	Month	Year	Center
MonthlyAll	73.1	37.1	93.6	79.7	49.7	26.1	57.9	86.5	98.0
All	84.5	30.7	96.7	77.0	48.3	13.8	60.4	79.4	86.9
ByCenter	91.2	33.7	98.3	76.0	52.5	23.1	79.1	92.0	71.2

corresponding to the plain ν in Table A1. These values are given in Table A2.

The reduction factors are essentially negligible (i.e., $\gamma_i \sim 1$) for domain, month, and year as expected, due to the nearly zero correlations for these dimensions. There is only a small reduction ($\gamma_i \sim 0.8$) for statistic, variable, and center. Reductions are more substantial for forecast time ($\gamma_i = 0.47$), level ($\gamma_i = 0.55$), and day ($\gamma_i = 0.66$).

The use of the specialized Monthly reference sample is critical in the calculation of the γ_i. That is because the effect of NWP center and trends with time induce large correlations that substantially reduce the ν_i and γ_i. Table A3 presents a sample of these calculations. Considering just the verification time dimensions, these γ_i correspond to one independent sample each 8, 15, and 6 days instead of one independent sample each 1.5 days for the Monthly reference sample used. The heuristic of applying Eq. (A8) to the "residuals" of the "model" of the NAMs as a function of NWP center and verification month is akin to the usual practice in statistical model building of examining residuals to see if there are trends or correlations or nonuniformity. In the present case the NAMs created with the specialized reference sample (Monthly) are the residuals, and by construction they have nearly uniform distributions. However, there are remaining correlations, and these are accounted for in the definition of n^*.

e. Acronyms

Key acronyms used in the text are listed here. Common acronyms (e.g., UTC and RMSE) and proper names (e.g., names of specific institutions such as NCEP and names of systems and version identifiers such as IFS and PS) are not expanded in the text.

3DEnVar Three-dimensional ensemble variational data assimilation
4DEnVar Four-dimensional ensemble variational data assimilation
AC Anomaly correlation
AME Absolute mean error
AR Autoregressive
CICS Cooperative Institute for Climate Studies (College Park, Maryland)
CIMAS Cooperative Institute for Marine and Atmospheric Studies (Miami, Florida)
ECDF Empirical cumulative density function
ECMWF European Centre for Medium-Range Weather Forecasts
EMC Environmental Modeling Center (NOAA/NWS)
GO General Operations
GRIB Gridded binary
IFS Integrated Forecast System
LCDNV Lead Centre for Deterministic NWP Verification
MA Moving average
NAM Normalized assessment metric
NCAR National Center for Atmospheric Research
NCEP National Centers for Environmental Prediction (NOAA/NWS)
NEMSIO NOAA Environmental Modeling System Input/Output (file format)
NHX Northern Hemisphere extratropics
NOAA National Oceanic and Atmospheric Administration
NWP Numerical weather prediction
NWS National Weather Service
OFS Overall forecast skill (score)
OSE Observing system experiment
OSSE Observing system simulation experiment
PAM Primary assessment metric
pdf Probability density function
PS Parallel Suite (UKMO)
RMSD Root-mean-square difference
RMSE Root-mean-square error
SAM Summary assessment metric
SCN Service change notice (NWS)
SHX Southern Hemisphere extratropics
TIN Technical implementation notice (NWS)
UKMO Met Office
USAF U.S. Air Force
UTC Coordinated universal time
VSDB Verification Statistics Database
WMO World Meteorological Organization (Geneva, Switzerland)

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, https://doi.org/10.1175/MWR-D-16-0013.1.

——, and Coauthors, 2018: Community Global Observing System Simulation Experiment (OSSE) Package (CGOP): Assessment and validation of the OSSE system using an OSSE–OSE intercomparison of summary assessment metrics. *J. Atmos. Oceanic Technol.*, **35**, 2061–2078, https://doi.org/10.1175/JTECH-D-18-0061.1.

Bretherton, C. S., M. Widmann, V. P. Dymnikov, J. M. Wallace, and I. Bladé, 1999: The effective number of spatial degrees of freedom of a time-varying field. *J. Climate*, **12**, 1990–2009, https://doi.org/10.1175/1520-0442(1999)012<1990:TENOSD>2.0.CO;2.

Buizza, R., G. Balsamo, and T. Haiden, 2018: IFS upgrade brings more seamless coupled forecasts. *ECMWF Newsletter*, No. 156, 18–22, ECMWF, Reading, United Kingdom, https://www.ecmwf.int/en/elibrary/14578-newsletter-no135-spring-2013.

Geer, A. J., 2016: Significance of changes in medium-range forecast scores. *Tellus*, **68A**, 30229, https://doi.org/10.3402/tellusa.v68.30229.

Hoffman, R. N., S.-A. Boukabara, V. K. Kumar, K. Garrett, S. P. F. Casey, and R. Atlas, 2017a: An empirical cumulative density function approach to defining summary NWP forecast assessment metrics. *Mon. Wea. Rev.*, **145**, 1427–1435, https://doi.org/10.1175/MWR-D-16-0271.1.

——, ——, ——, ——, ——, and ——, 2017b: A non-parametric definition of summary NWP forecast assessment metrics. *28th Conf. on Weather Analysis and Forecasting/24th Conf. on Numerical Weather Prediction*, Seattle, WA, Amer. Meteor. Soc., 618, https://ams.confex.com/ams/97Annual/webprogram/Paper309748.html.

Janousek, M., 2018: Score definitions and requirements. WMO Lead Centre for Deterministic NWP Verification, ECMWF, https://software.ecmwf.int/wiki/display/WLD/Score+definitions+and+requirements.

Lorenz, E. N., 1982: Atmospheric predictability experiments with a large numerical model. *Tellus*, **34**, 505–513, https://doi.org/10.3402/tellusa.v34i6.10836.

Newman, K. M., 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*, Austin, TX, Amer. Meteor. Soc., 620, 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, https://doi.org/10.1002/qj.32.

Shafran, P., T. L. Jensen, J. H. Gotway, B. Zhou, K. Nevins, Y. Lin, and G. DiMego, 2015: Web-based verification capability using NCEP's verification database and DTC's METviewer. *27th Conf. on Hurricanes and Tropical Meteorology*, Chicago, IL, Amer. Meteor. Soc., 14A.7, http://ams.confex.com/ams/27WAF23NWP/webprogram/Paper273777.html.

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

Zhou, B., and Coauthors, 2015: An overview of grid-to-grid verification at Environmental Modeling Center (EMC) of NCEP. *27th Conf. on Hurricanes and Tropical Meteorology*, Chicago, IL, Amer. Meteor. Soc., 12A.4, http://ams.confex.com/ams/27WAF23NWP/webprogram/Paper273363.html.