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

SID-AHMED BOUKABARA,[a] KAYO IDE,[b] YAN ZHOU,[c,a] NARGES SHAHROUDI,[d,a] ROSS N. HOFFMAN,[e,f] KEVIN GARRETT,[a] V. KRISHNA KUMAR,[d,a] TONG ZHU,[g,a] AND ROBERT ATLAS[f]

[a] *NOAA/NESDIS/Center for Satellite Applications and Research (STAR), College Park, Maryland*
[b] *University of Maryland, College Park, College Park, Maryland*
[c] *Cooperative Institute for Climate and Satellites, University of Maryland, College Park, College Park, Maryland*
[d] *Riverside Technology Inc., College Park, Maryland*
[e] *Cooperative Institute for Marine and Atmospheric Studies, University of Miami, Miami, Florida*
[f] *NOAA/Atlantic Oceanographic and Meteorological Laboratory, Miami, Florida*
[g] *Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, Colorado*

(Manuscript received 9 April 2018, in final form 6 June 2018)

ABSTRACT

Observing system simulation experiments (OSSEs) are used to simulate and assess the impacts of new observing systems planned for the future or the impacts of adopting new techniques for exploiting data or for forecasting. This study focuses on the impacts of satellite data on global numerical weather prediction (NWP) systems. Since OSSEs are based on simulations of nature and observations, reliable results require that the OSSE system be validated. This validation involves cycles of assessment and calibration of the individual system components, as well as the complete system, with the end goal of reproducing the behavior of real-data observing system experiments (OSEs). This study investigates the accuracy of the calibration of an OSSE system—here, the Community Global OSSE Package (CGOP) system—before any explicit tuning has been performed by performing an intercomparison of the OSSE summary assessment metrics (SAMs) with those obtained from parallel real-data OSEs. The main conclusion reached in this study is that, based on the SAMs, the CGOP is able to reproduce aspects of the analysis and forecast performance of parallel OSEs despite the simplifications employed in the OSSEs. This conclusion holds even when the SAMs are stratified by various subsets (the tropics only, temperature only, etc.).

1. Introduction and study objectives

Observing system simulation experiments (OSSEs[1]) allow ''what if'' experiments that quantify the expected real-world impact of changes to observing systems—the focus of this study—or changes to data assimilation (DA) and forecast systems. OSSEs are based on simulations of nature and observations, and compare results from a control configuration and a test configuration. The *control* configuration (usually) assimilates all currently available observations into a DA and forecast system that is as close as possible to existing operational practice. In the *test* configuration, either a new observing system is added to the control configuration or the DA and forecast system is modified relative to the control configuration. By way of example and introduction: The nature run (NR) of the OSSE system used here is a 2-yr-long 7-km-resolution forecast of the Goddard Earth Observing System Model, version 5 (GEOS-5). This NR is commonly referred to as the GEOS-5 nature run (G5NR) (Putman et al. 2015). The OSSE system—the Community Global OSSE Package (CGOP) (Boukabara et al. 2016b)—contains several components: First, the CGOP interpolates the G5NR in time and space, and applies observation operators for conventional, radiance, and radio occultation (RO) observations (Boukabara et al. 2018). Then, the CGOP includes a research version of the NOAA global DA system (GDAS) that assimilates these observations. A goal of the OSSE assimilation step is to faithfully reproduce the effects of the quality

[1] All acronyms are defined in the appendix.

Corresponding author: Dr. Sid-Ahmed Boukabara, sid.boukabara@noaa.gov

DOI: 10.1175/JTECH-D-18-0061.1

control, data selection, and assimilation methodology. Finally, the CGOP includes verification and visualization tools to assess and interpret the impact results. Here, we compare CGOP results to parallel real-data observing system experiments (OSEs) of Boukabara et al. (2016a, hereafter BGK).

Results from OSSEs are used 1) to inform decision-making processes designed to optimize global orbital configurations, 2) in sensor design trade studies, 3) as a test bed to prototype and implement new techniques, and 4) to increase readiness to exploit new observing systems prior to their actual deployment. The design of an OSSE should be tailored to the particular goals of the OSSE (Hoffman and Atlas 2016). Ideally, the DA and forecast systems used for the simulated-data OSSEs and the real-data OSEs should be close to the operational system.

Since OSSEs are based on simulations of nature and observations, reliable results require that the OSSE system be validated. A validated system is capable of reproducing the behavior of real-data OSEs. The validation process normally includes cycles of assessment and calibration (i.e., tuning) of the individual system components as well as the complete system, by comparing OSSE and OSE results. The first CGOP component to be validated was the simulation of error-free observations (Boukabara et al. 2018), which have then been used in subsequent OSSEs, including those reported here.

In the present study, we assess the calibration of the entire CGOP before any explicit tuning has been performed by conducting an OSSE–OSE intercomparison in a parallel setting. This component of the OSSE system validation examines to what extent the simulation-based OSSE results agree with real-data OSE results, for similar observing system configurations. We expect the OSSE results—when examining assessment metric by assessment metric with no normalization—to be better than the OSE results for two reasons. First, the OSSE forecast model in the CGOP is likely more similar to the NR model than the OSE forecast model is to reality. Second, and specific to the OSSEs used here, the observations are *perfect* observations without any added explicit errors. Thus, such OSSE results are expected to fail an absolute assessment; that is, primary assessment metrics (PAMs) such as the 500-hPa geopotential height anomaly correlation (AC) or the 250-hPa wind RMSE are expected to be quite different in the OSSE compared to the OSE. However, OSSE results may still be useful for some purposes if the relative impacts are reliable in the sense of being equivalent to the relative impacts observed in the OSEs. Thus, if after appropriate differencing, scaling, or normalization the adjusted assessment metrics agree, then the OSSE results have passed a relative assessment test. This may be sufficient for assessing the impact of a degraded or improved constellation of observing systems or the impact of replacing sensors or satellites with alternative ones.

Some previous studies have sought to validate and calibrate OSSEs (e.g., Errico et al. 2013; Privé et al. 2014). For validation, Hoffman and Atlas (2016) list the following steps. First, deficiencies of the NR should be assessed and documented. Any experiment using the NR must examine whether the NR deficiencies would conflict with the requirements of that experiment or invalidate the assumptions of that experiment. Second, the DA and forecast error statistics should be similar for impact tests that can be conducted in both OSSE and reality, such as those examined here. Two additional recommended tests are performed to validate the predictability characteristics of the OSSE system compared to reality and to compare OSSEs and OSEs at the start of the NR before it diverges from reality.

Significant differences found in the validation may require calibrating the OSSE system by adjusting the simulated observation errors to better match relevant statistics, usually statistics of the observation innovations [observation minus background ($O - B$)] (e.g., Errico et al. 2013) or modifying some of the forecast model parameterizations in the OSSE to better match the forecast skill in reality (e.g., Casey et al. 2015). Unfortunately, tuning system parameters, which might be the estimated observation error statistics used by the assimilation or the ''constants'' appearing in the model parameterizations of small-scale physical processes, for the purpose of calibration may result in values that are unphysical or inconsistent with prior knowledge. Moreover, operational DA systems commonly apply bias-correction schemes (e.g., Dee and Uppala 2009) to the observation innovations, which may interact with adjustment and tuning processes. A significant challenge to traditional calibration is that focusing on an individual primary assessment metric (PAM; i.e., a particular statistic for an given forecast time, domain, variable, and level) may not help and may actually harm the degree of agreement for other PAMs. This is one reason to consider summary assessment metrics (SAMs) when validating and calibrating an OSSE system. As an alternative to tuning the OSSE system to match reality, OSSE results can be directly calibrated to comparable OSE results. For example, Hoffman et al. [1990, section 5c(2)] tuned OSSE impacts by assuming OSE impacts are proportional to OSSE impacts and determining that ratio from the parallel OSSE and OSE results for withholding satellite observations.

The context and plan of this study are as follows: The CGOP, including the simulation of and caveats related to perfect observations, is described in section 2.

This section also briefly describes those OSEs conducted by BGK to explore the impact of reductions on satellite observations that are used here as well as the parallel OSSEs conducted for this study. The current study continues the validation of the CGOP by comparing OSSE and OSE SAMs. A focus on SAMs increases statistical significance and avoids the problems of focusing on one particular PAM, as discussed above. Here, SAMs are averages of normalized assessment metrics (NAMs) and each NAM corresponds to a single PAM. For example, the PAM for the 120-h 500-hPa geopotential height Northern Hemisphere (NH) forecast AC valid 8 August 2006 for the control OSSE is converted to the corresponding NAM using a normalization that is based on a reference sample of the 120-h 500-hPa geopotential height NH forecast AC PAMs for all valid verification times and experiments.[2] While a first-order validation might consider ''global'' SAMs that combine all NAMs for each experiment, subsequent validation might consider SAMs for various subsets or along different dimensions (e.g., for each forecast time or domain). Originally, in what we termed ''overall'' scores, a minimum–maximum (minmax) normalization was used (e.g., BGK). Later, we proposed the alternative use of an empirical cumulative density function (ECDF) normalization and applied it to the OSEs of BGK (Hoffman et al. 2017b) and to the 2015 skill scores from several global NWP centers (Hoffman et al. 2017a).[3] The minmax and ECDF SAMs are described and contrasted in section 3. Also in section 3, the effect of using the control analyses (i.e., the analyses from the control experiments) for verification is investigated by comparison to using the NR for verification for the OSSEs. Comparisons of OSSEs to OSEs in section 4 first show 500-hPa AC and 250-hPa wind RMSE as typical PAMs. As expected the OSSE results for this idealized configuration of the CGOP are superior. However, OSSE and OSE results appear to be more similar in terms of scorecards of impact significance, and in terms of maps and cross sections of errors[4] after these have been scaled by the domainwide magnitude of the errors to account for the difference in the magnitudes of the errors of the OSSEs and OSEs. ECDF and minmax SAMs are then compared along several dimensions. We find that focusing on SAMs represents a useful approach for validating idealized OSSEs. Striking similarities of the results between OSSE and OSE are found using both ECDF and minmax SAMs. Finally, section 5 provides a discussion and concluding remarks, including a discussion of the relationship to absolute OSSE system calibration.

2. OSSE methodology

a. The CGOP

The CGOP is an evolving package as described in detail by Boukabara et al. (2016b). In brief the CGOP includes the G5NR developed by NASA (Putman et al. 2015); forward operators to simulate error-free observations—including the Community Radiative Transfer Model (CRTM; Chen et al. 2008; Ding et al. 2011) and the RO observation simulator developed by NOAA (Cucurull et al. 2013); an observation error addition procedure developed by NASA (Errico et al. 2013); and a DA system—the operational hybrid 3D- and 4D-ensemble variational (3DEnVar and 4DEnVar, respectively) GDAS developed by NOAA (Kleist and Ide 2015a,b).

The G5NR is a global atmospheric 7-km nonhydrostatic forecast from 16 May 2005 until 16 June 2007, forced by the observed sea surface temperature. NASA conducted extensive validation of the G5NR in comparison to reality (Gelaro et al. 2015). The G5NR is a very detailed simulation, including representations of extreme weather events (e.g., Reale et al. 2017). However, there are significant differences compared to reality as documented by Gelaro et al. (2015) and some that may be important in the OSSE context are listed by Boukabara et al. (2018). For example, because of diffusion added for computational stability, the effective resolution (Skamarock 2004) of the G5NR should be considered as several times coarser than its 7-km grid spacing (Gelaro et al. 2015, section 2.2).

The observation operators, described in brief by Boukabara et al. (2018), are the same for simulating observations for the OSSE and within the GDAS with exceptions noted in what follows. The microwave and infrared brightness temperatures (BTs) are simulated using the CRTM (Chen et al. 2008; Ding et al. 2011) and RO refractivities, and bending angles are simulated following Cucurull et al. (2013). Although these same observation operators are used in simulating observations from the G5NR and within the DA system, the preliminary spatial and temporal interpolation to the observation locations is not identical, since the spatial grids and time archiving are different. In general, but not in the present idealized study, errors are tuned and explicitly added to the simulated perfect observations following the method of Errico et al. (2013).

[2] Other SAMs, such as the NWP index of Rawlins et al. (2007), use skill scores to normalize PAMs and weighted averages to define SAMs.

[3] A new study by Hoffman et al. (2018) explores this topic in more detail.

[4] Here, what is termed error is really an approximate error, since forecasts are compared to analyses that are imperfect. Only in those results where the OSSEs are verified against the NR are the ''errors'' true, not approximate errors.

The CGOP includes the NOAA GDAS and forecast systems—the operational hybrid 3DEnVar and 4DEnVar GDAS with 80 ensemble members (Kleist and Ide 2015a,b). The current operational GDAS configuration (see NWS 2014) uses a 64-layer sigma–pressure hybrid coordinate, T1534 resolution for the deterministic forecast, T574 resolution for the 80-member forecast ensemble, and T574 resolution for the 4DEnVar analysis. The research version uses the same 64 vertical layers, but T670 resolution for the deterministic forecast and T254 resolution for the forecast ensemble used in the 3DEnVar or 4DEnVar. Note that the resolution of these components is given in terms of their spectral truncation. For example, T574 means triangular truncation at total wavenumber 574. See Table 3 in Boukabara et al. (2016b) for conversions to kilometers.

b. *Experiment setup*

Results shown below (sections 3 and 4) are for three of the four OSEs described by BGK and for three parallel OSSEs. These experiments examine two 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 GDAS, which then included the hybrid 3DEnVar. Of the four experiments conducted by BGK, the three experiments considered here are as follows:

- cntrl-OSE: All observing systems used in the January 2015 operational implementation are included in the control configuration in this best-case experiment.
- 3polar-OSE: This experiment reduces satellite observation coverage to the 3-Polar configuration in which all secondary and backup polar satellites are eliminated, thereby retaining only one satellite in each of the early morning, midmorning, and evening orbits.
- 2polar-OSE: This experiment further reduces satellite observation coverage to the 2-Polar configuration, which is the same as the 3-Polar configuration but without the evening platform.

Each experiment covered the period 25 May–7 August 2014, but only the 32-day period of 7 July–7 August 2014 was used for intercomparison purposes. Forecasts were made each day at 0000 UTC for 168 h (7 days). (For the purpose of this paper, we have added the ''OSE'' suffix to the experiment names of BGK.)

In summary, using minmax SAMs, BGK (p. 2547) finds that ''removing secondary satellites results in significant degradation of the forecast. Second, losing the afternoon orbit on top of losing secondary satellites further degrades forecast performances by a significant margin.'' These findings are consistent with the results presented here using a new assessment metric, that is, using ECDF SAMs.

For intercomparison, parallel OSSEs were conducted. These experiments are similarly named, but with ''OSSE'' as a suffix—cntrl-OSSE, 3polar-OSSE, and 2polar-OSSE. In what follows, the cntrl-, 3polar- and 2polar-OSSEs and OSEs will be referred to in pairs as the Control, 3-Polar, and 2-Polar experiments. The 32-day OSSE period, 15 August–15 September 2006, follows a 7-day spinup period, 8–15 August 2006. The OSSEs were run using the research version of the GDAS. To minimize configuration differences between the OSEs of BGK and the parallel OSSEs, 1) the GDAS uses 3DEnVar in this study and 2) data in the OSSE are simulated to match the actual 2014 observation locations and times (i.e., the months, days, and times of day shifted to 2006). Thus, the OSSEs and OSEs are identical in experimental procedures with the following exceptions. First, the OSSE resolutions are T670 and T254 and the OSE resolutions are T1534 and T574. Second, there are differences in the data sampling because of the difference in time of year. While the exact locations of observations were different during 7 July–7 August 2014 and during 15 August–15 September 2014, the approximate numbers and coverage for each data type were the same.

3. SAM methodology

Various assessment metrics can be compared between OSSEs and OSEs to establish the degree of similarity between the OSSE and reality. In what follows, the focus is on the behavior of SAMs for the three OSEs and the three parallel OSSEs described in section 2b. For the SAMs presented in what follows, along each PAM dimension (e.g., variable) there are a number of discrete coordinate values (geopotential height, temperature, etc.) that are listed here in ''dimension: coordinate values'' format:

1) forecast time: 0, 24, 48, 72, 96, 120, 144, 168 h;
2) level: 250, 500, 700, 850, 1000 hPa;
3) domain: Northern Hemisphere extratropics (NHX), Southern Hemisphere extratropics (SHX), tropics;
4) variable: geopotential height Z, temperature T, vector wind $\mathbf{V}$, relative humidity (RH);
5) statistic: anomaly correlation (AC), root-mean-square error (RMSE), absolute mean error (AME);
6) verification time: the 32 OSSE or OSE verification times, which occur every 24 h at 0000 UTC; and
7) experiment: {cntrl-OSSE, 3polar-OSSE, 2polar-OSSE} or {cntrl-OSE, 3polar-OSE, 2polar-OSE}.

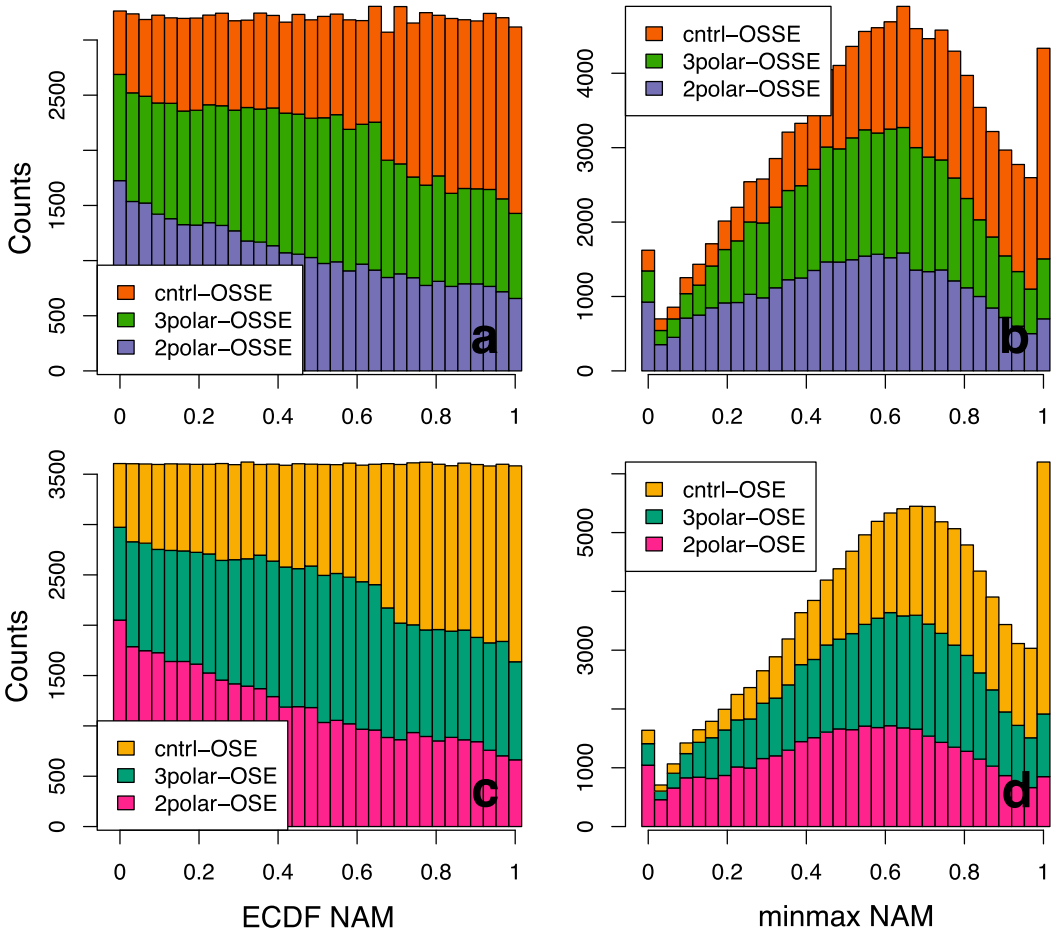

FIG. 1. Stacked histograms for (a),(b) OSSEs and (c),(d) OSEs of (left) ECDF NAMs and (right) minmax NAMs. The total number of NAMs (and of PAMs) is 102 600 for the OSSEs and 115 200 for the OSEs. The distributions are strikingly similar for both ECDF and minmax normalizations.

The PAMs are calculated with respect to the corresponding Control analysis—either cntrl-OSSE or cntrl-OSE—and the SAMs are calculated using the ECDF normalization. Note that for the vector wind PAMs, ordinary multiplications are replaced with dot products in the definitions of the statistical quantities. Alternatives to these approaches—using the minmax normalization and verifying against the NR—are discussed here, and some results using these alternatives are presented and/or discussed. The available OSE PAMs for both RMSE and AME are complete, including all possible 46 080 combinations of coordinates. Some of the possible AC PAMs are not calculated in keeping with standard practice. As a result there are no RH AC PAMs, and ACs are missing for 850-hPa geopotential height, and for 700- and 1000-hPa temperature and wind, leaving a total of 23 040 OSE AC PAMs. For the OSSEs, for the first few days there are missing values, since the forecasts start with the first verification day (e.g., there is no 3-day forecast present at day 2 of the OSSEs), leaving a total of 41 040 RMSE and AME PAMs, and 20 520 AC PAMs.

In prior research, we have calculated SAMs using both ECDF and minmax normalizations (Hoffman

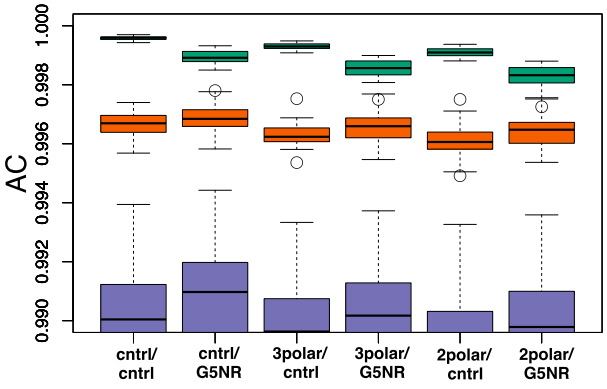

FIG. 2. Boxplots of PAMs for 500-hPa geopotential height AC at forecast time 0 (mint), 24 (ochre), and 48 h (lavender) for the OSSEs using both Control-verification (cntrl) and NR-verification (G5NR).

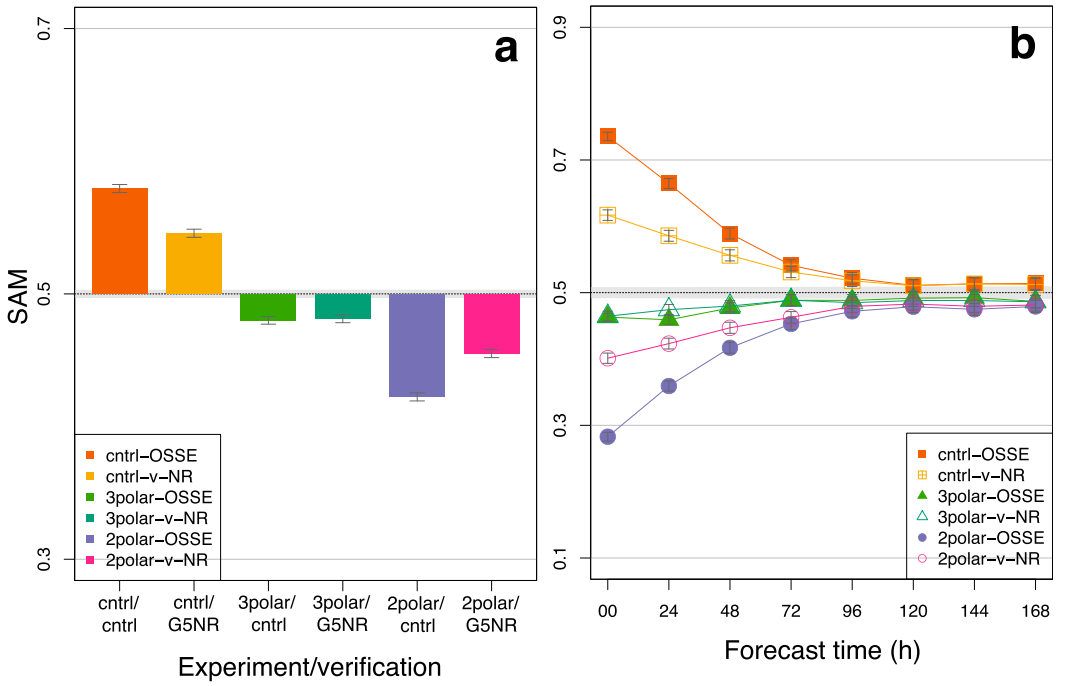

FIG. 3. Effect of Control- vs NR-verification on (a) global SAMs and (b) SAMs as a function of forecast time using ECDF normalization for the OSSEs. Note that in this and subsequent plots of SAM, confidence intervals are plotted at the 95% level, gray shading indicates the 95% null hypothesis (H_0) confidence interval, and the y axis ranges from 0.3 to 0.7, or from 0.1 to 0.9 in cases when the x axis is forecast time.

et al. (2017b; BGK). In both cases the normalization is specific to each PAM subset, that is, each individual forecast time, level, domain, variable, and statistic (e.g., all the 120-h 500-hPa geopotential height NHX forecast ACs). For each PAM subset, the reference sample $\mathcal{R}$ includes the PAMs for all verification times and all experiments in either all OSSEs or all OSEs. In the present case, there are 1440 PAM subsets (eight forecast times, five levels, three domains, four variables, and three statistics), and the maximum reference sample size is 96 (32 verification times and three experiments). For a PAM like AC, where increasing values are better, the ECDF normalization is given by

$$\mathrm{NAM} = \frac{\mathrm{Rank}(\mathrm{PAM\ in}\ \mathcal{R}) - 1}{\mathrm{Size}(\mathcal{R})}, \tag{1}$$

and the minmax normalization is given by

$$\mathrm{NAM} = \frac{\mathrm{PAM} - \min(\mathcal{R})}{\max(\mathcal{R}) - \min(\mathcal{R})}. \tag{2}$$

In both formulations NAMs are in the range [0, 1], with 0 being worst and 1 best. For a PAM like RMSE, where increasing values are worse, Eqs. (1) and (2) are applied to the negative of the PAM values.

In general, the ECDF normalization may be preferred because of its robustness and well-behaved known distribution, while the minmax normalization may be preferred because of its simple implementation. In practice, some artifacts occur when using either normalization. Figure 1 shows that the shape of the ECDF and minmax distributions are quite different.[5] Consider the 2polar-OSSE experiments. The ECDF distribution decreases approximately linearly with NAM value, while the minmax distribution is quasi-Gaussian plus a constant. The distributions for OSSEs and OSEs are very similar—a much stronger statement than that the means of the distributions are similar. Under the conditions of this experiment, the OSSE results even for perfect observations are consistent with the real-data OSE results. By construction the combined (stacked) histogram using ECDF must be uniform. This is not the case for the minmax normalization, where there is a well-defined mode centered on 0.65. This distribution is asymmetric and is consistent with the fact that there are occasional forecast busts so that the minimum will be

[5] Colors in this and other figures are from the Dark2 color-blind safe palette of ColorBrewer—mint, ochre, lavender, magenta, lime, gold, and brown. Many of the following figures use the colors scheme of Fig. 1 to indicate the different OSSEs and OSEs.

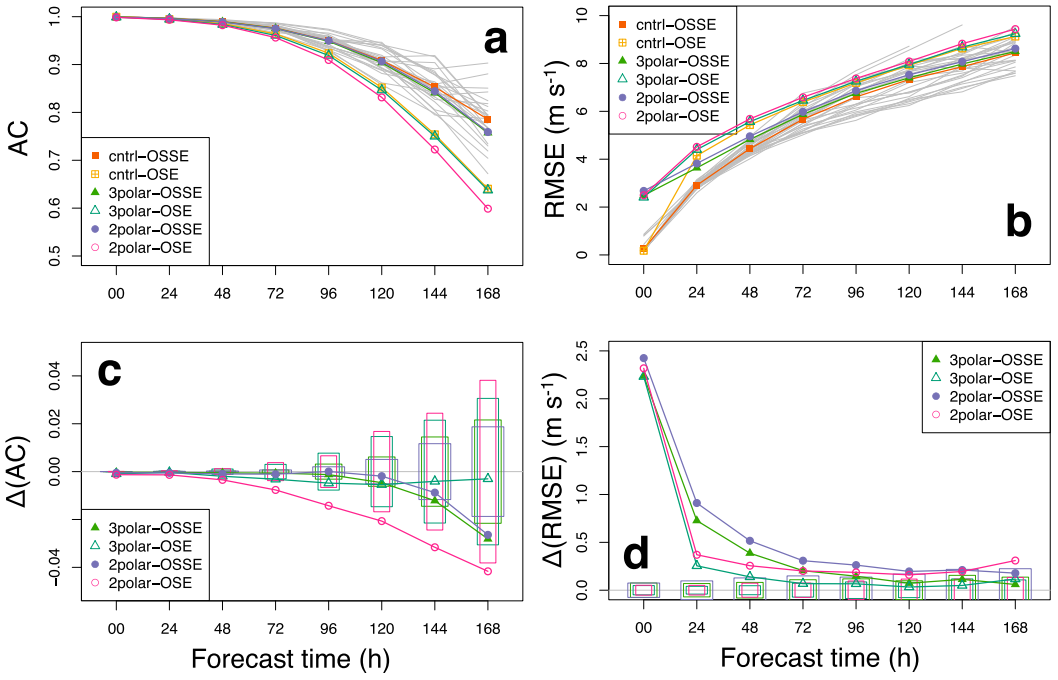

FIG. 4. Plots of (a) AC of 500-hPa geopotential height in the NHX and (b) RMSE of 250-hPa wind in the tropics vs forecast time. The PAMs are plotted for each cntrl-OSSE forecast in gray under the average curves. (c),(d) Differences with respect to each Control experiment and 95% significance intervals (boxes) for the null hypotheses.

more extreme than the maximum. The accumulation of NAMs at 0 and 1 using the minmax normalization is an artifact of the minmax calculation—there must be one occurrence of 0 and one occurrence of 1 for each of the 1440 subsets (i.e., for each $\mathcal{R}$). The ECDF NAM values are quantized (from 0 to 95, divided by 95). In the histograms plotted, each bin includes exactly three of these quantized values. For the OSEs there are no missing values and small deviations from flatness are due to ties in the ranks. Since there are missing values at the start of the OSSEs, small-amplitude random noise was added to the OSSE NAM values to stabilize this histogram (Fig. 1a). For the minmax normalization, since the values are not quantized, there is no need to stabilize the histograms and the histograms are relatively insensitive to the choice of binning.

The OSSE assessment metrics may be calculated with verification from the Control experiment of the NR (Control- or NR-verification). This has a large effect at the initial (0 h) forecast time but a decreasing effect thereafter. This is clearly seen in Figs. 2 and 3. Figure 2 show the effect of Control- versus NR-verification at the start of the forecast, zooming in on just the 0-, 24-, and 48-h 500-hPa NHX geopotential height AC. At 0 h, the Control experiments verified against the corresponding Control analyses are nearly perfect, since the Control 0-h forecast is a one-time-step forecast from the analysis. For all cases, at 0 h, the verification against cntrl-OSSE is slightly better than against the G5NR, but the reverse is true at 24 and 48 h.

Hoffman et al. (2017b) found that the global SAM (i.e., the average of all NAMs) is a useful summary of skill and that there is a decay of impact with forecast time. In 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 as a result of the experiment. Thus, ECDF SAM values of 0.75 and 0.25 would represent very large positive and negative impacts, respectively, since the expected value under a null hypothesis H_0 is 0.5. Figure 3 examines how the choice of verification affects the key SAM characteristics. In Fig. 3a the global SAMs are compared for the OSSEs for ECDF normalization, for both Control- and NR-verification. Figure 3 and subsequent figures identify SAM results for the OSSEs verified against the G5NR as cntrl-v-NR, 3polar-v-NR, and 2polar-v-NR, whereas all SAM results verified against the Control analyses are simply identified by the experiment name. For reasons best understood in terms of Fig. 3b, the impacts are somewhat smaller using the NR as verification. In Fig. 3b the SAMs are plotted versus forecast time. The effect of using Control-verification is to increase Control skill at 0 h. This effect decays with forecast time and is near zero by 72 h. In the ECDF case, the

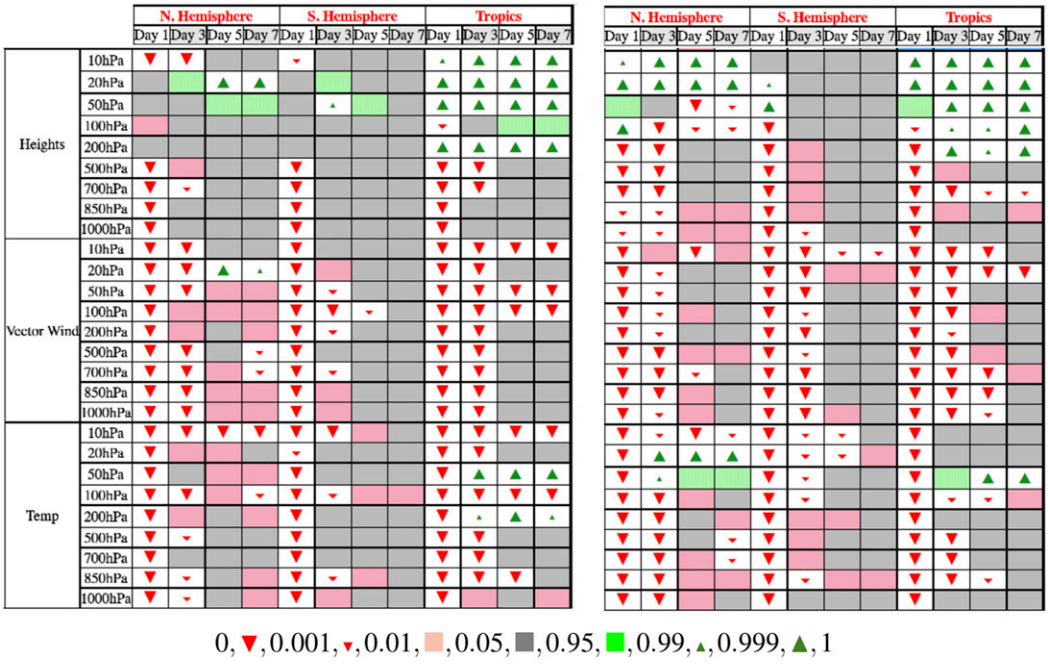

FIG. 5. Scorecards for the 3-Polar experiments compared to the Control experiments for the (left) OSSEs and (right) OSEs. Only the RMSE portions of the standard NOAA Environmental Modeling Center verification scorecards are shown. The symbols and colors indicate the probability that the 3-Polar experiment is better than the Control experiment. As shown below the scorecard, the green symbols indicate that the 3-Polar experiment is better than the Control experiment at the 95% (squares), 99% (small up-pointing triangle), and 99.9% (large up-pointing triangle) significance levels, while the red symbols indicate that the 3-Polar experiment is worse than the Control experiment at the 95% (squares), 99% (small down-pointing triangle), and 99.9% (large down-pointing triangle) significance levels. Gray indicates no statistically significant differences.

effect of verification on cntrl-OSSE must be offset by the sum of the effects on the other experiments. For the present experiments, all of this effect is distributed to 2polar-OSSE and there is little effect on 3polar-OSSE. Since there is no absolute truth for the OSEs, for consistency in those plots that compare OSSE and OSE results, assessment metrics will be calculated with Control-verification.

4. OSSE–OSE intercomparison results

a. PAMs

Examples of standard ''loss of predictability'' curves for PAMs using Control-verification are shown in Fig. 4. The perfect-observations OSSEs have higher predictability than the real-data OSEs as expected. For cntrl-OSSE, the OSSE AC reach 0.8 at 168 h, while the OSE AC reach 0.8 at about 130 h. The impact of the 2-Polar configuration compared to Control is much greater and more significant in the OSE than in the OSSE setting. The underlying plume plot in Fig. 4a shows each of the cntrl-OSSE AC curves. Clearly there is substantial variability hidden in each average curve and that variability grows with forecast time. The RMSE plot shows similar behavior. However, here it is easy to see that at 0 h the Control experiments are nearly perfect and much better than the other experiments as a result of using the Control analyses for verification (as mentioned earlier, these are not exactly perfect because of small differences between the Control analyses and 0-h forecasts, which are one-time-step forecasts from the analysis). Beyond 48 h, the wind forecast errors grow nearly linearly with time through the end of the 7-day forecast. Note that the later forecast time OSSE results are approximately 1 day better (e.g., the 5-day OSSE scores are as good as the 4-day OSE scores).

The curves in the bottom (difference) panels in Fig. 4 agree somewhat better than in the top panels, but there are noticeable differences. All the data gap experiments result in poorer forecast scores, but for the geopotential height AC, 2polar-OSE is definitely worse than 2polar-OSSE. For the wind RMSE, all the results are similar except at 24 and 48 h, where the OSSE forecasts have larger impacts. The 95%-significance-level uncertainties support the robustness of these results. Note that in Fig. 4c, 2polar-OSSE is slightly better than 3polar-OSSE.

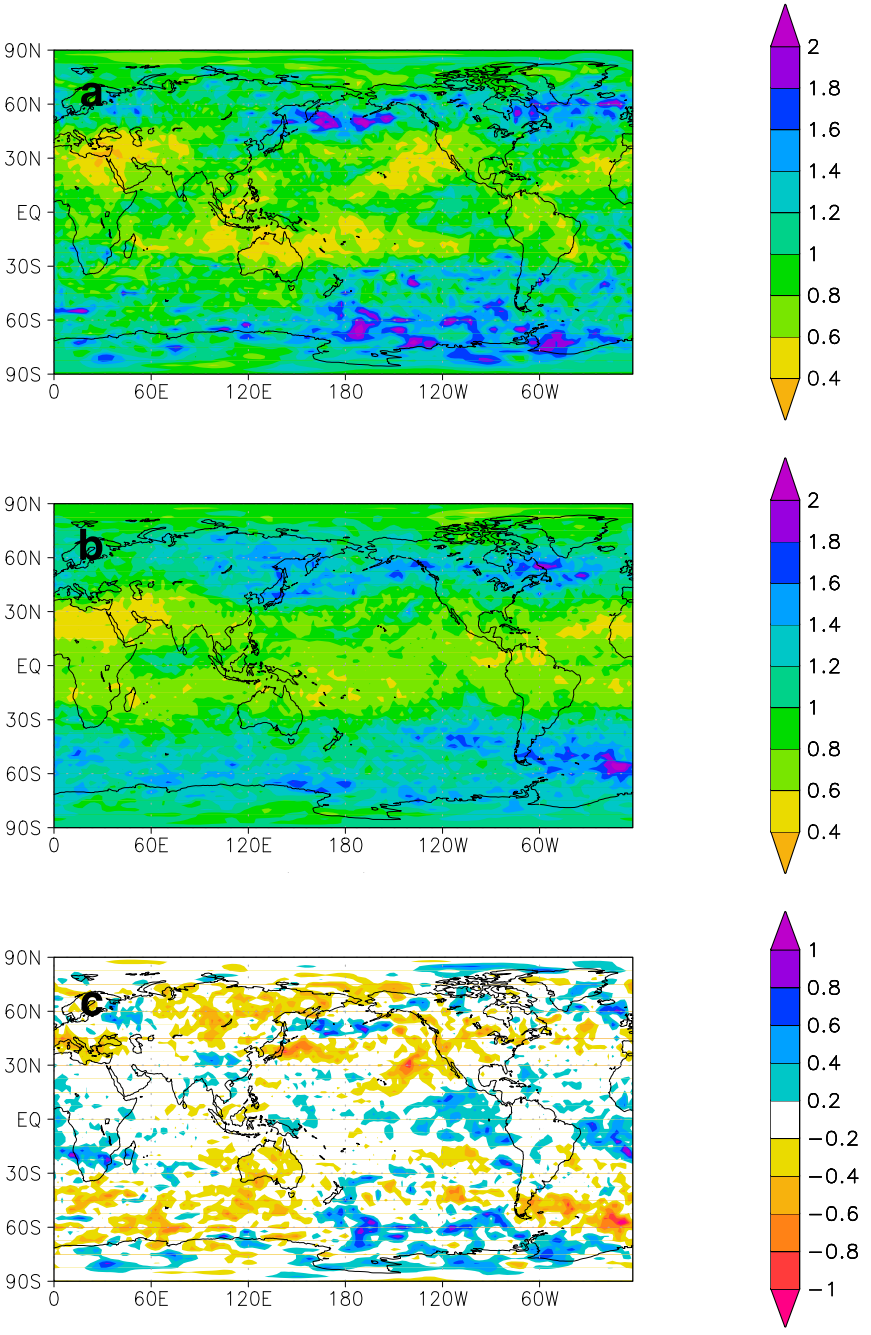

FIG. 6. Maps of 72-h 250-hPa wind-forecast-error-scaled standard deviation for (a) the 3polar-OSSE, (b) the 3polar-OSE, and (c) the difference (OSSE minus OSE). The standard deviation is over all forecasts, and the scales are the RMS of the OSSE standard deviations (4.46 m s^{-1}) and the RMS of the OSE standard deviations (5.51 m s^{-1}).

Such nonintuitive results within the estimated uncertainty bounds are common when examining individual PAMs. This is one motivation to use SAMs.

In figures like Fig. 4 only a small fraction of the PAMs can be displayed. Therefore, it is common to collectively visualize many impacts in a scorecard, as in Fig. 5, which compares the 3-Polar experiments to the Control experiments. There is substantial agreement between OSSE and OSE in terms of which impacts of the 3-Polar configuration are significantly negative and positive and to what degree of significance. Note that in both the OSSE and OSE, the 3-Polar configuration generally

degrades forecast skill, but not for geopotential height and some temperatures in the extratropical stratosphere (as with the uncertainty boxes in the bottom panels in Fig. 4, each symbol in Fig. 5 corresponds to an individual two-sided paired Student's *t* test).

The results presented so far in this section are for statistics computed over large domains, but such visualizations hide the substantial spatial variation. Figure 6 shows maps of the 72-h 250-hPa scaled standard deviation of the wind forecast error for 3polar-OSSE (top) and 3polar-OSE (middle). Since the overall magnitude of the errors is considerably smaller in the OSSE, scaling both the OSSE and OSE maps by their RMS values helps to assess to what extent the patterns of errors agree. The overall patterns in Figs. 6a and 6b agree fairly well on the large scale, but there are many small-scale differences (Fig. 6c). This might be expected, since the individual daily error maps that are averaged are dominated by errors on the synoptic scale, and the OSSE and OSE synoptic features do not match. Note that the wind errors tend to be largest where the winds are strongest (60°N/S) and smallest where the winds are weakest (the Middle East).

Visualizations like Fig. 6 show only one level and one forecast time, so vertical cross sections are used to examine other features of the F7 forecast errors. Figure 7 displays cross sections of 120-h geopotential height zonal-mean scaled forecast error standard deviation for 3polar-OSSE (Fig. 7a) and 3polar-OSE (Fig. 7b). The largest errors are in the jet stream regions, where there can be mislocations of the traveling synoptic-scale waves. The patterns of the scaled errors in the OSSE and OSE agree fairly well. Again, the scale of the errors is considerably smaller in the OSSE, with the OSSE RMS standard deviation equal to 82% of the OSE value.

Similar conclusions are drawn from examining maps like those in Fig. 6 for different levels, variables, and forecast lengths, and cross sections like those in Fig. 7 for different variables and forecast lengths. Generally, the largest differences occur where the scaled errors are largest (e.g., Fig. 7c between 60° and 90°S), and the scaled errors tend to be largest in the extratropics and near the tropopause.

The findings of this section that simple differencing or scaling transformations reveal similarities in the OSSE and OSE results motivate the use of SAMs for assessing the OSSE system calibration and validation.

b. SAMs

As noted earlier, Hoffman et al. (2017b) found that the key SAM characteristics for data impact experiments are that global SAMs are a useful summary of skill and that there is a clear decay of impact with forecast time. For the OSSE–OSE intercomparison, Fig. 8a plots the global SAMs, that is, NAMs averaged over all dimensions—statistic, forecast time, variable, level, and domain. As an aid to visualization, the OSSE and OSE results are plotted pairwise by observing configuration (this figure is in the same format as Fig. 3, but with the OSE results replacing the NR-verification OSSE results). There is very good agreement between OSSE and OSE global SAMs—in fact, much better than the comparison between Control- and NR-verification OSSE results seen in Fig. 3a.

In data impact experiments, SAMs vary greatly with forecast time. Figure 8b shows the quasi-exponential decay of impact with forecast time for these experiments.

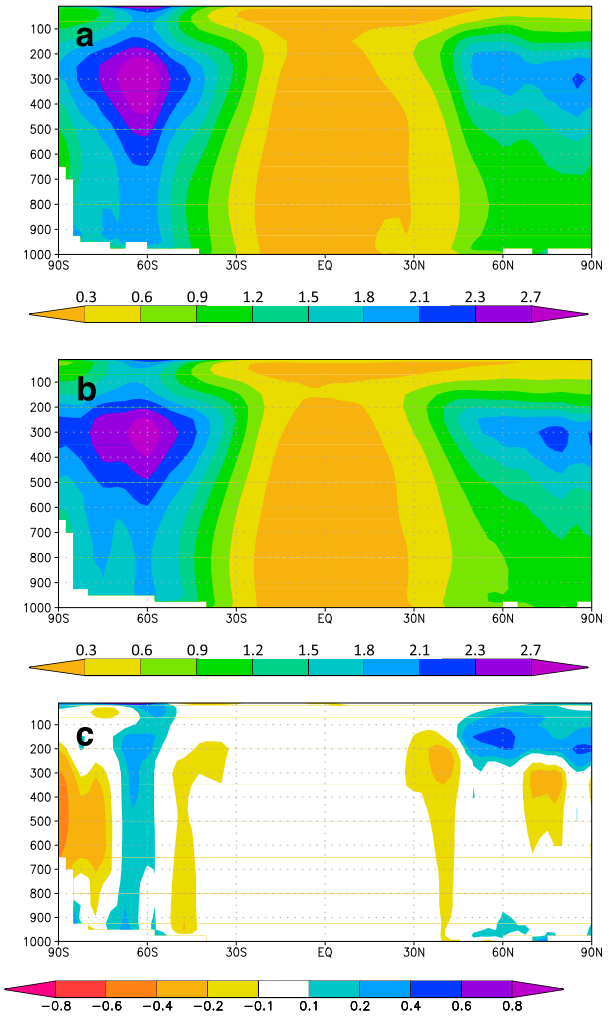

FIG. 7. Cross sections of 120-h geopotential height zonal-mean forecast-error-scaled standard deviation for (a) the 3polar-OSSE, (b) the 3polar-OSE, and (c) the difference (OSSE minus OSE). The standard deviation is over all longitudes and then averaged over all forecasts, and the scales are the RMS of the OSSE standard deviations (31.5 m) and the RMS of the OSE standard deviations (38.2 m).

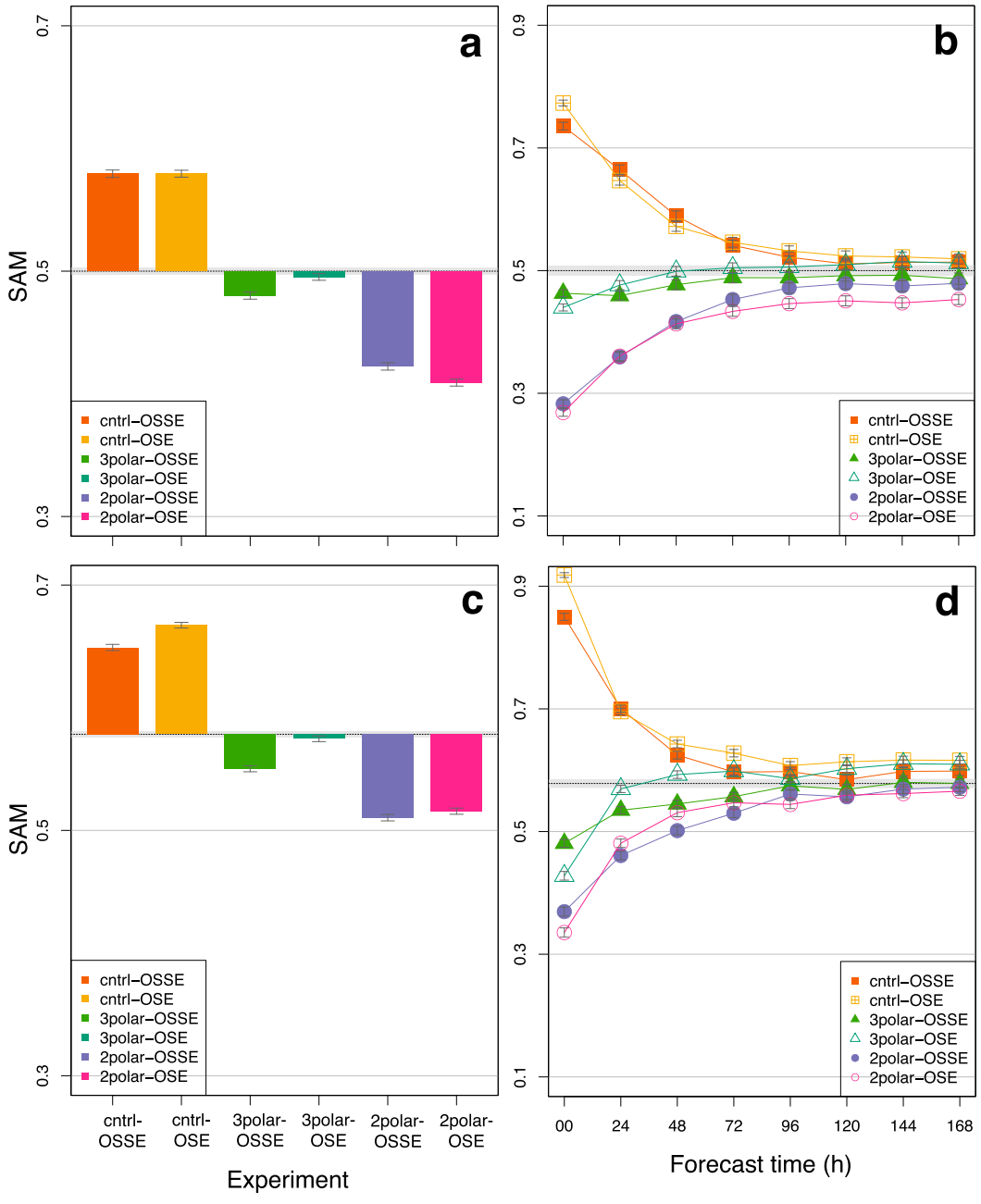

FIG. 8. (a) Global ECDF SAMs and (b) ECDF SAMs as a function of forecast time for OSSEs and OSEs. (c),(d) As in (a) and (b), respectively, but showing the minmax SAM results.

At longer forecast times, the negative impact of the 2polar-OSSE (lavender filled circle) is smaller than that of the 2polar-OSE (magenta open circle), but for both 3polar-OSSE (lime filled triangle) and 3polar-OSE (mint open triangle) the impacts are nearly neutral, although of opposite sign. This strong similarity between OSSE and OSE SAMs is confirmed as well using the minmax normalization. Generally similar results are found for global SAMs using the minmax normalization (Fig. 8c). The general variations with forecast time are similar, but the negative impact seen in 2polar-OSSE is not noticeably reduced in magnitude compared to 2polar-OSE, and at 0 h there are larger impacts for the Control and 3-Polar experiments (Fig. 8d).

Figure 9 shows how SAMs vary along the dimensions of domain, variable, level, and statistic. Across all categories the OSSE and OSE impacts are very

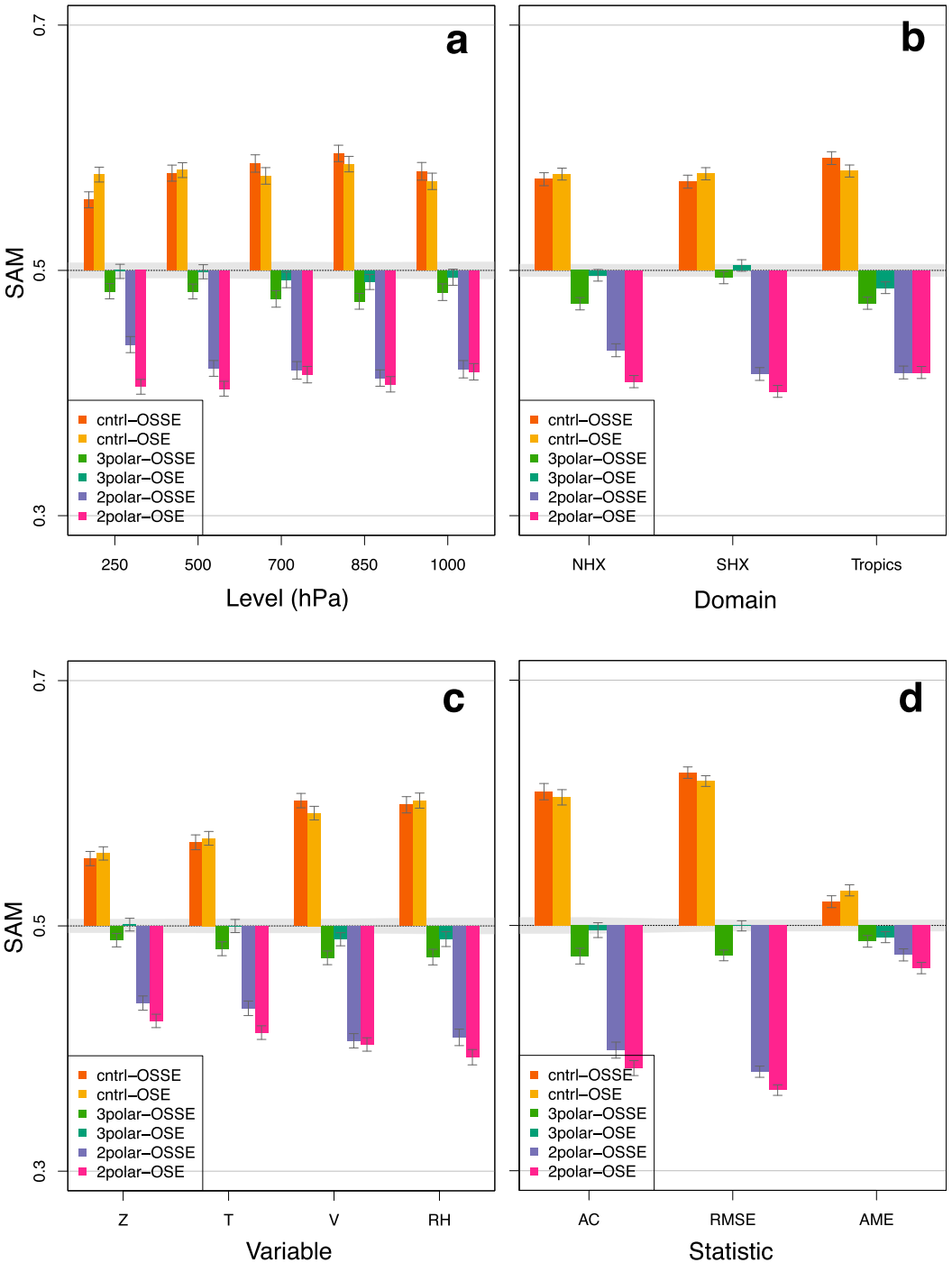

FIG. 9. ECDF SAMs for OSSEs and OSEs by (a) level (hPa), (b) domain, (c) variable, and (d) statistic.

similar for the Control experiments, 2polar-OSSE tends to have smaller negative impacts than 2polar-OSE, and the 3polar-OSSE impacts are generally small and negative, while the 3polar-OSE impacts are mostly neutral. Notably, impacts for wind and RH are larger than for geopotential height and temperature, and impacts for AME are much smaller than for AC and RMSE. The OSSE and OSE SAM impacts for geopotential height now agree unlike the PAM impacts shown in Fig. 4. The striking similarity of these plots suggests that for SAMs, the OSSE results are consistent with the OSE results. Here also, the strong similarity between OSSE

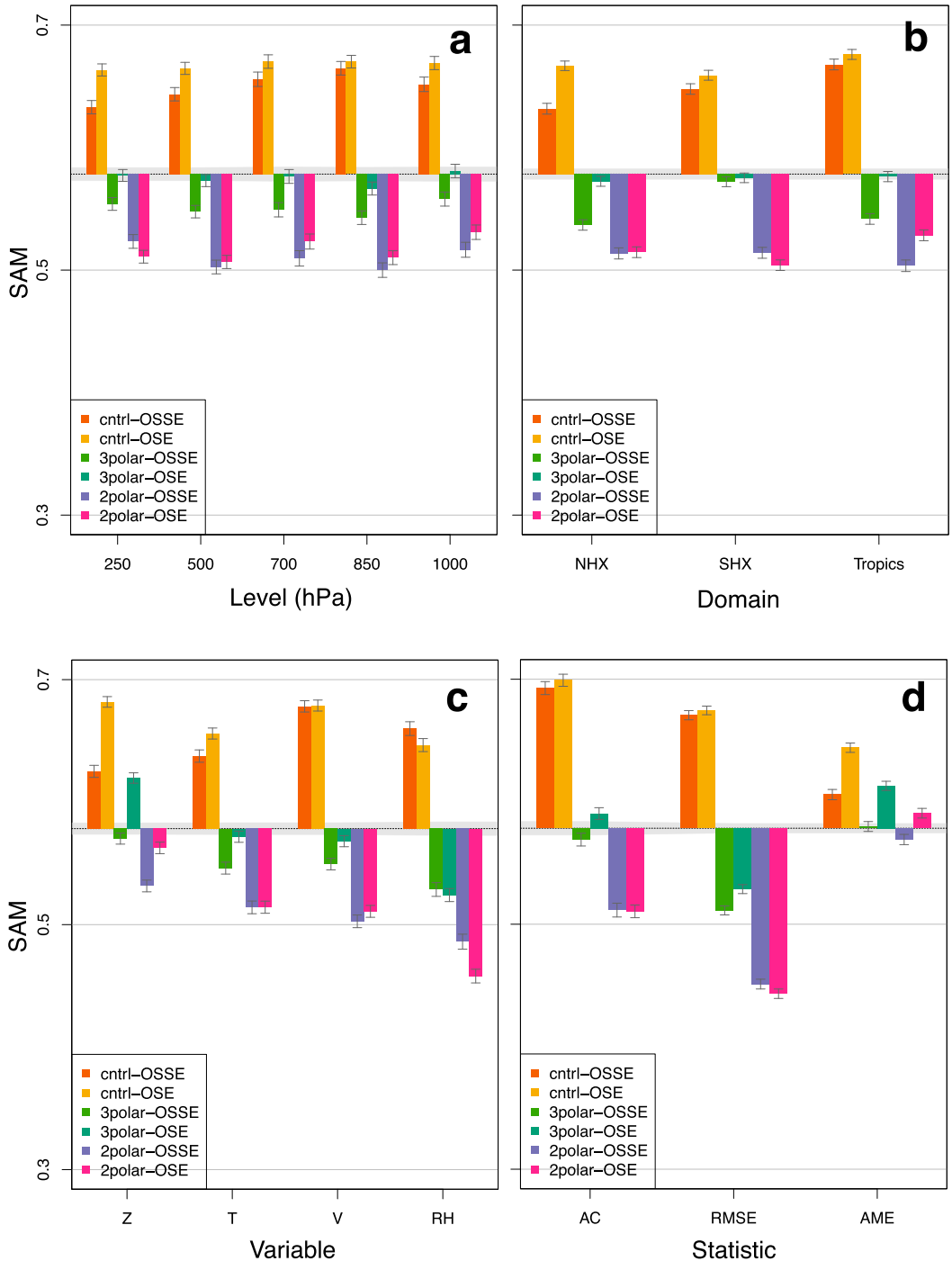

FIG. 10. Minmax SAMs for OSSEs and OSEs for (a) level (hPa), (b) domain, (c) variable, and (d) statistic.

and OSE SAMs is found using the minmax normalization. Conclusions from ECDF and minmax normalization are similar (Fig. 10), except that ECDF suggests a stronger similarity for some categories, including for the NH domain, and for geopotential height. (The results in Fig. 10d appear somewhat inconsistent because of the use of a single baseline for the bar plot. The mean value of the minmax NAMs actually varies noticeably between AC, RMSE, and AME.) Comparing the OSSE impacts for Control- and NR-verification (Fig. 11), the impacts are generally reduced using NR-verification (everything is closer to 0.5) as expected from the discussion in

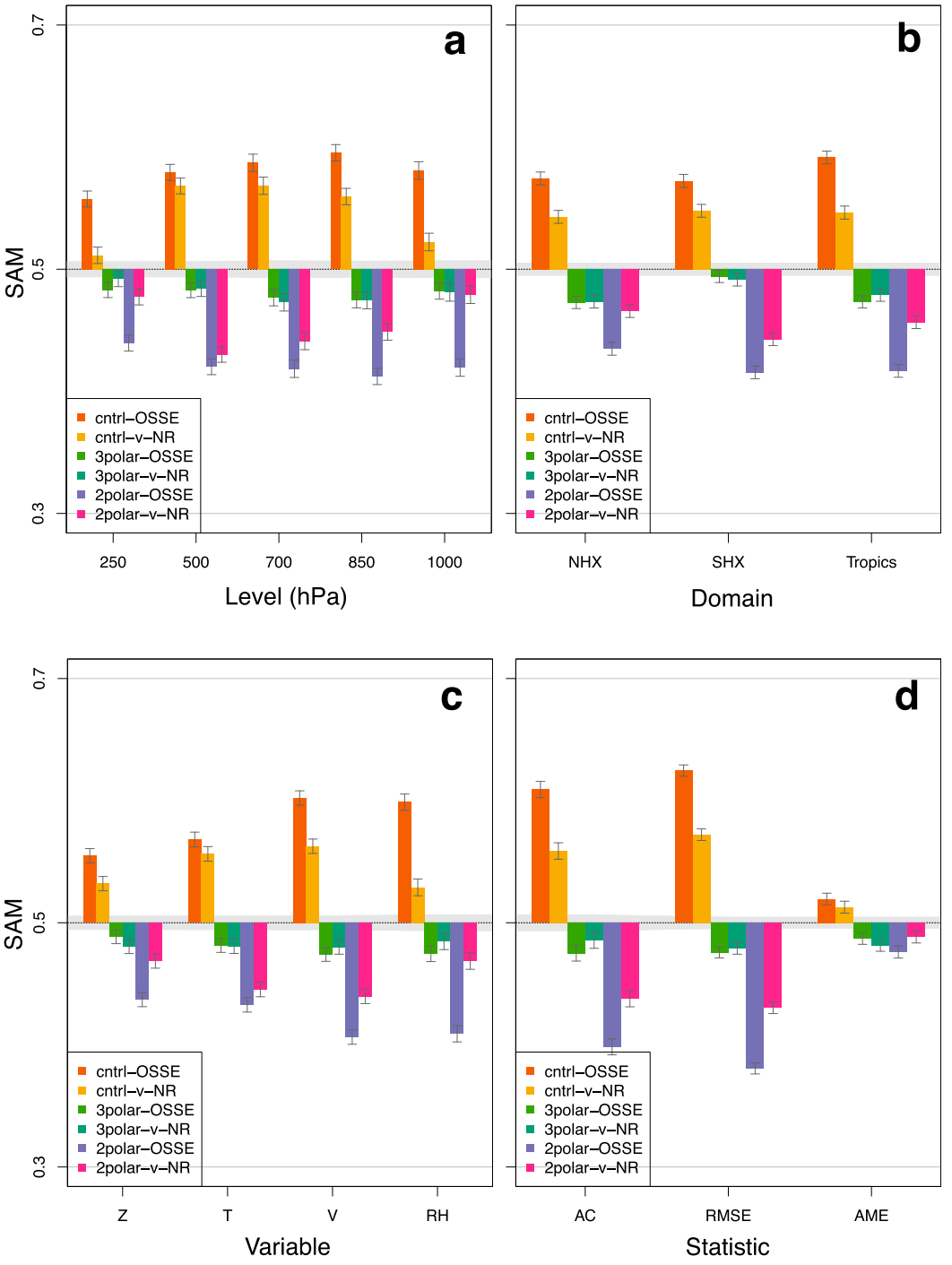

FIG. 11. ECDF SAMs for OSSEs for Control- and NR-verification for (a) level (hPa), (b) domain, (c) variable, and (d) statistic.

section 3. But the NR-verification impacts are even more reduced in certain categories, including RH compared to other variables and at the highest and lowest levels (250 and 1000 hPa, respectively) compared to midlevels. This indicates that in these categories the Control analysis is less reliable and that it might make more sense to verify the OSSEs against the NR and the OSEs against the operational consensus analysis as has been done in previous OSSEs (e.g., Atlas 1997; Atlas et al. 1985a,b, 2001, 2015a,b).

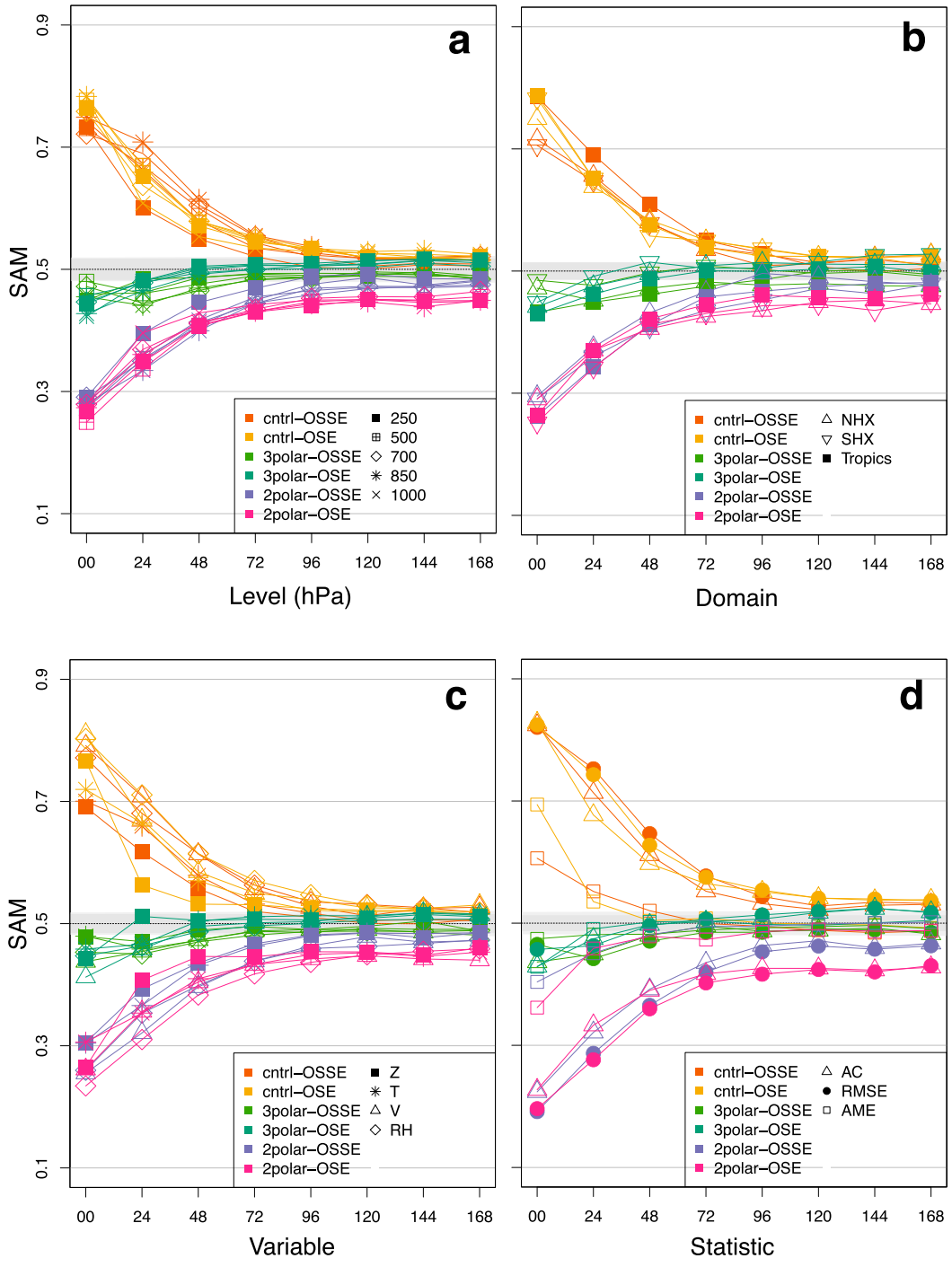

FIG. 12. ECDF SAMs for OSSEs and OSEs as functions of forecast time by (a) level (hPa), (b) domain, (c) variable, and (d) statistic.

Further decomposition of SAMs by forecast time and other dimensions shows generally similar patterns in the OSSE and OSE results (Fig. 12). Essentially the interaction of forecast time and the other dimensions (of domain, variable, level, and statistic) are weak and these decompositions are cross products of Fig. 8b and the individual panels of Fig. 9. One notable variation (Fig. 12c) is that relative to the other variables, the cntrl-OSE geopotential height impacts are reduced in magnitude at 24 and 48 h.

5. Discussion and concluding remarks

OSSEs must be realistic and reliable to accurately predict the impact of future observing systems or changes to DA systems, and to support decision-making. The relevance of any OSSE results to specific conclusions must be considered in light of the deficiencies of the particular OSSE system that was used. In an ideal situation, the OSSE system uses the operational DA system, and the OSSE is validated by showing that the OSSE impacts accurately reflect the OSE impacts for experiments that can be executed in parallel, as well as the absolute accuracies of analyses and forecasts. Often, however, the validation assessment in terms of PAMs tends to suggest the need to calibrate or tune the OSSE system or to calibrate the conclusions of the experiment based on the limitations of the OSSE system that is used. Calibration of the OSSE system might include adjusting the simulated observation error characteristics (Errico et al. 2013; Privé et al. 2014) or modifying some parameters in the forecast model to affect its predictability characteristics. However, using single metrics of single parameters for the calibration assessment is challenging (i.e., no OSSE calibration to date has been found to be able to reproduce real-data performance for every single metric of every single parameter). A posteriori calibration of the results determines adjustments that make the parallel OSSEs and OSEs similar and then applies the same adjustments to the other OSSEs.

This present study suggests a method of assessing the OSSE system calibration based on an OSSE–OSE intercomparison of SAMs. This study focused on the validation of the OSSE impacts of potential data gaps on global operational NWP using the CGOP. The results indicate that this OSSE system even without calibration or tuning can be useful in evaluating at least one of the important questions typically addressed in an OSSE—specifically that of determining the relative impact of observing systems. Ongoing work is underway to repeat the OSSEs with realistic added observation errors (following Errico et al. 2013). The present study is also limited to experiments for the impacts of groups of sensors; in the future, additional experiments similar to those described here could be conducted examining the impacts of individual sensors.

OSSEs with perfect observations cannot by definition explore the impact of varying observation error magnitudes and correlations. This constitutes a limitation to the application of such idealized OSSEs. It is however worth noting that tuning observation errors has its own challenges as described earlier. Furthermore, when assessing new sensors, real error characteristics are often unknown and therefore any attempt to adjust errors of future sensors may run the risk of underestimating or overestimating their actual values. It is important that baseline sensor-simulated observations and new sensor-simulated observations be treated consistently. While acknowledging the aforementioned limitations, it should also be noted that OSSEs with perfect observations can be used, similar to how the ensemble of DA approach (e.g., Harnisch et al. 2013) has been used, to investigate the impact of the distribution of observations and what parameters are observed at those locations. An appropriate validation for perfect observation OSSEs is then to show that the OSSE impacts agree with the OSE impacts across variables and locations in a relative or normalized sense; that is, some characteristics of the impacts should be similar, and simple transformations (differencing, scaling, normalizing) can provide a first step in validation and calibration.

Acknowledgments. The authors thank the many colleagues and collaborating scientists who contributed by their interactions, peer-reviews, and suggestions, including John Pereira, Tim Schmit, Jun Li, Joanne Ostroy, Kevin Tewey, Frank Gallagher, David Spencer and Karen St. Germain.

We gratefully acknowledge support for this work provided by NOAA—particularly by the NOAA National Environmental Satellite, Data, and Information Service (NESDIS) Office of Projects, Planning, and Analysis (OPPA); and the NOAA Office of Systems Architecture and Advance Planning (OSAAP), under the Scientific and Technical Services II Federal Contracting Vehicle (DOCDG133E12CQ0020), and the auspices of the cooperative institutes given in the author affiliations on the title page through Cooperative Agreements NA14NES4320003 for CICS, NA15OAR4320064 for CIMAS, and NA14OAR4320125 for CIRA.

APPENDIX

Acronyms

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 NASA) are not expanded in the text.

3DEnVar	3D-ensemble variational
4DEnVar	4D-ensemble variational
AC	Anomaly correlation
AME	Absolute mean error
BGK	Boukabara et al. (2016a)
BT	Brightness temperature

CGOP	Community Global OSSE Package
CICS	Cooperative Institute for Climate and Satellites (College Park, Maryland)
CIMAS	Cooperative Institute for Marine and Atmospheric Studies (Miami, Florida)
CIRA	Cooperative Institute for Research in the Atmosphere (Fort Collins, Colorado)
CRTM	Community Radiative Transfer Model
DA	Data assimilation
ECDF	Empirical cumulative density function
G5NR	GEOS-5 nature run [GMAO 7-km (1/16° × 1/16°)-resolution NR]
GDAS	Global DA system
GEOS-5	Goddard Earth Observing System Model, version 5 (NASA)
GMAO	Global Modeling and Assimilation Office
GOS	Global observing system
H0	Null hypothesis
NAM	Normalized assessment metric
NASA	National Aeronautics and Space Administration
NESDIS	National Environmental Satellite, Data, and Information Service
NH	Northern Hemisphere
NHX	Northern Hemisphere extratropics
NOAA	National Oceanic and Atmospheric Administration
NR	Nature run
NWP	Numerical weather prediction
OPPA	Office of Projects, Planning, and Analysis
OSAAP	Office of Systems Architecture and Advance Planning
OSE	Observing system experiments
OSSE	Observing system simulation experiment
PAM	Primary assessment metrics
RH	Relative humidity
RMSE	Root-mean-square error
RO	Radio occultation
SAM	Summary assessment metric
SHX	Southern Hemisphere extratropics
UTC	Coordinated universal time

REFERENCES

Atlas, R., 1997: Atmospheric observations and experiments to assess their usefulness in data assimilation. *J. Meteor. Soc. Japan*, **75**, 111–130, https://doi.org/10.2151/jmsj1965.75.1B_111.

——, E. Kalnay, W. E. Baker, J. Susskind, D. Reuter, and M. Halem, 1985a: Simulation studies of the impact of future observing systems on weather prediction. Preprints, *Seventh Conf. on Numerical Weather Prediction*, Montreal, QC, Canada, Amer. Meteor. Soc., 145–151.

——, ——, and M. Halem, 1985b: Impact of satellite temperature soundings and wind data on numerical weather prediction. *Opt. Eng.*, **24**, 242341, https://doi.org/10.1117/12.7973481.

——, and Coauthors, 2001: The effects of marine winds from scatterometer data on weather analysis and forecasting. *Bull. Amer. Meteor. Soc.*, **82**, 1965–1990, https://doi.org/10.1175/1520-0477(2001)082<1965:TEOMWF>2.3.CO;2.

——, L. Bucci, B. Annane, R. Hoffman, and S. Murillo, 2015a: Observing system simulation experiments to assess the potential impact of new observing systems on hurricane forecasting. *Mar. Technol. Soc. J.*, **49**, 140–148, https://doi.org/10.4031/MTSJ.49.6.3.

——, and Coauthors, 2015b: Observing system simulation experiments (OSSEs) to evaluate the potential impact of an optical autocovariance wind lidar (OAWL) on numerical weather prediction. *J. Atmos. Oceanic Technol.*, **32**, 1593–1613, https://doi.org/10.1175/JTECH-D-15-0038.1.

Boukabara, S.-A., K. Garrett, and V. K. Kumar, 2016a: 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, 2016b: Community Global Observing System Simulation Experiment (OSSE) Package (CGOP): Description and usage. *J. Atmos. Oceanic Technol.*, **33**, 1759–1777, https://doi.org/10.1175/JTECH-D-16-0012.1.

——, and Coauthors, 2018: Community Global Observing System Simulation Experiment (OSSE) Package (CGOP): Perfect observations simulation validation. *J. Atmos. Oceanic Technol.*, **35**, 207–226, https://doi.org/10.1175/JTECH-D-17-0077.1.

Casey, S. P. F., H. Wang, R. Atlas, R. N. Hoffman, S.-A. Boukabara, Y. Xie, Z. Toth, and J. S. Woollen, 2015: Initial validation of a new OSSE capability. *19th Conf. on Integrated Observing and Assimilation Systems for the Atmosphere, Oceans, and Land Surface (IOAS-AOLS)*, Phoenix, AZ, Amer. Meteor. Soc., 3.2, https://ams.confex.com/ams/95Annual/webprogram/Paper267725.html.

Chen, Y., F. Weng, Y. Han, and Q. Liu, 2008: Validation of the Community Radiative Transfer Model by using CloudSat data. *J. Geophys. Res.*, **113**, 2156–2202, https://doi.org/10.1029/2007JD009561.

Cucurull, L., J. C. Derber, and R. J. Purser, 2013: A bending angle forward operator for global positioning system radio occultation measurements. *J. Geophys. Res. Atmos.*, **118**, 14–28, https://doi.org/10.1029/2012JD017782.

Dee, D. P., and S. Uppala, 2009: Variational bias correction of satellite radiance data in the ERA-Interim reanalysis. *Quart. J. Roy. Meteor. Soc.*, **135**, 1830–1841, https://doi.org/10.1002/qj.493.

Ding, S., P. Yang, F. Weng, Q. Liu, Y. Han, P. van Delst, J. Li, and B. Baum, 2011: Validation of the Community Radiative Transfer Model. *J. Quant. Spectrosc. Radiat. Transfer*, **112**, 1050–1064, https://doi.org/10.1016/j.jqsrt.2010.11.009.

Errico, R. M., R. Yang, N. C. Privé, K.-S. Tai, R. Todling, M. E. Sienkiewicz, and J. Guo, 2013: Development and validation of observing-system simulation experiments at NASA's Global Modeling and Assimilation Office. *Quart. J. Roy. Meteor. Soc.*, **139**, 1162–1178, https://doi.org/10.1002/qj.2027.

Gelaro, R., and Coauthors, 2015: Evaluation of the 7-km GEOS-5 nature run. Technical Report Series on Global Modeling and Data Assimilation, Vol. 36, R. D. Koster, Ed., NASA Tech.

Memo. TM-2014-104606v36, 285 pp., http://gmao.gsfc.nasa.gov/pubs/docs/Gelaro736.pdf.

Harnisch, F., S. B. Healy, P. Bauer, and S. J. English, 2013: Scaling of GNSS radio occultation impact with observation number using an ensemble of data assimilations. *Mon. Wea. Rev.*, **141**, 4395–4413, https://doi.org/10.1175/MWR-D-13-00098.1.

Hoffman, R. N., and R. Atlas, 2016: Future observing system simulation experiments. *Bull. Amer. Meteor. Soc.*, **97**, 1601–1616, https://doi.org/10.1175/BAMS-D-15-00200.1.

——, C. Grassotti, R. G. Isaacs, J.-F. Louis, and T. Nehrkorn, 1990: Assessment of the impact of simulated satellite lidar wind and retrieved 183 GHz water vapor observations on a global data assimilation system. *Mon. Wea. Rev.*, **118**, 2513–2542, https://doi.org/10.1175/1520-0493(1990)118<2513:AOTIOS>2.0.CO;2.

——, S.-A. Boukabara, V. K. Kumar, K. Garrett, S. P. F. Casey, and R. Atlas, 2017a: 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.

——, ——, ——, ——, ——, and ——, 2017b: 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.

——, V. K. Kumar, S.-A. Boukabara, K. Ide, F. Yang, and R. Atlas, 2018: Progress in forecast skill at three leading global operational NWP centers during 2015–2017 as seen in Summary Assessment Metrics (SAMs). *Wea. Forecasting*, accepted, https://doi.org/10.1175/WAF-D-18-0117.1, in press.

Kleist, D. T., and K. Ide, 2015a: An OSSE-based evaluation of hybrid variational–ensemble data assimilation for the NCEP GFS. Part I: System description and 3D-Hybrid results. *Mon. Wea. Rev.*, **143**, 433–451, https://doi.org/10.1175/MWR-D-13-00351.1.

——, and ——, 2015b: An OSSE-based evaluation of hybrid variational–ensemble data assimilation for the NCEP GFS. Part II: 4DEnVar and hybrid variants. *Mon. Wea. Rev.*, **143**, 452–470, https://doi.org/10.1175/MWR-D-13-00350.1.

NWS, 2014: Corrected: Global Forecast Systems (GFS) update: Effective January 14, 2015. National Weather Service Technical Implementation Notice 14-46, http://www.nws.noaa.gov/om/notification/tin14-46gfs_cca.htm.

Privé, N. C., Y. Xie, S. Koch, R. Atlas, S. J. Majumdar, and R. Hoffman, 2014: An observation system simulation experiment for the unmanned aircraft system data impact on tropical cyclone track forecasts. *Mon. Wea. Rev.*, **142**, 4357–4363, https://doi.org/10.1175/MWR-D-14-00197.1.

Putman, W. M., A. Darmenov, A. da Silva, R. Gelaro, A. Molod, L. Ott, and M. J. Suarez, 2015: A 7-km non-hydrostatic global mesoscale simulation for OSSEs with the Goddard Earth Observing System Model (GEOS-5). *19th Conf. on Integrated Observing and Assimilation Systems for the Atmosphere, Oceans, and Land Surface (IOAS-AOLS)*, Phoenix, AZ, Amer. Meteor. Soc., 3.1, https://ams.confex.com/ams/95Annual/webprogram/Paper260701.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.

Reale, O., D. Achuthavarier, M. Fuentes, W. M. Putman, and G. Partyka, 2017: Tropical cyclones in the 7-km NASA global nature run for use in observing system simulation experiments. *J. Atmos. Oceanic Technol.*, **34**, 73–100, https://doi.org/10.1175/JTECH-D-16-0094.1.

Skamarock, W. C., 2004: Evaluating mesoscale NWP models using kinetic energy spectra. *Mon. Wea. Rev.*, **132**, 3019–3032, https://doi.org/10.1175/MWR2830.1.