Some General and Fundamental Requirements for Designing Observing System Simulation Experiments (OSSEs)

A White Paper prepared by
Ronald M. Errico[1] and Nikki C. Privé

Goddard Earth Sciences, Technology and Research Center, Morgan State University
and
Global Modeling and Assimilation Office, NASA Goddard Space Flight Center

Submitted 18 April 2018
Revised 3 August 2018

Executive Summary

The intent of this white paper is to inform WMO projects and working groups, together with the broader weather research and general meteorology and oceanography communities, regarding the use of Observing System Simulation Experiments (OSSEs). This paper is not intended to be either a critical or cursory review of past OSSE efforts. Instead, it describes some fundamental, but often neglected, aspects of OSSEs and prescribes important caveats regarding their design, validation, and application.

Well designed, properly validated, and carefully conducted OSSEs can be invaluable for estimating and understanding impacts of proposed observing systems and new data assimilation techniques. Although significant imperfections and limitations should be expected, OSSEs either profoundly complement or uniquely provide both qualitative and quantitative characterizations of potential analysis of components of the earth system.

Goals: It is our intention to encourage the WMO and broader community to: (a) promote OSSE development and applications; (b) thoughtfully design OSSE frameworks; (c) carefully conduct appropriate experiments; (d) warn creators and users of OSSE results regarding sometimes poor practices, misinterpretations, or misrepresentations offered.

Challenges: (a) Since OSSEs are fundamentally a simulation of the very complex, interdisciplinary, statistical, and economically critical problem of data assimilation, they require extensive validation of all components before being applied. (b) Procurement, deployment, and implementation requirements demand that OSSE guidance and assessments be provided in a timely manner, creating constraints on the OSSE development and validation. (c) With plug and play or other community software available, temptations to conduct OSSEs with inadequate understanding of the components, algorithms, and limitations of either the simulations or their real data counterparts must be

[1] Corresponding author address: Dr. Ronald M. Errico, Global Modeling and Assimilation Office, Code 610.1, Goddard Space Flight Center, Greenbelt, MD 20771 USA; e-mail: ronald.m.errico@nasa.gov

discouraged. (d) With the primary impact of system errors (e.g., instrumentation, representation, algorithmic, and forecast model errors) on system results, adequate estimation and simulation of all error components must be addressed, especially correlated observation errors and forecast model errors.

Recommendations: The WMO and broader community should (a) provide thoughtful input and encourage development of OSSE nature runs at willing and able institutions intended for wide ranges of applications by the community, as well as assist in required, extensive validation efforts; (b) provide educational opportunities for would-be developers and practitioners, especially for those lacking knowledge or experience; (c) encourage collaborative development and free exchanges of OSSE components, if not of actual software, at least of their algorithms; (d) encourage greater consideration of the characterization of all types of system errors, qualitatively and quantitatively; (e) create formal opportunities for more significant critical reviews of OSSE frameworks and applications not only because they are generally intended to be scientifically based but because their results will potentially impact the entire community as the observing systems for the future are shaped.

1. Introduction

Observing System Simulation Experiments (OSSEs) are applications of data assimilation systems (DAS) conducted in an entirely simulated environment for estimating impacts of proposed data inputs or assimilation algorithms. Rather than using real observations of the real atmosphere that have real errors, OSSEs employ simulated observations of simulated atmospheres that have simulated errors. For this reason, they are not restricted to assimilating only existing observation types but can incorporate future envisioned ones as well. Also, unlike the real atmosphere, for which the true state is never known precisely, the simulated atmosphere that provides the "truth" for the analysis is known perfectly in an OSSE. These two properties render OSSEs extremely useful for a variety of applications.

When developing a new observing system, many competing designs usually present themselves, with each having its own cost and benefit. In an OSSE, simulated data having the planned spatial and temporal distributions and error characteristics of selected system designs can be assimilated along with simulated data representing other expected observations. The resulting data assimilation products then can be examined with regard to each particular design's impacts. In this way, OSSEs can estimate some of the expected benefits of each design or system before funds are expended actually manufacturing and deploying the instruments. Once a flexible, validated OSSE framework is available, little cost or time is required to conduct the required comparative investigations. This traditionally has been the motivation for OSSEs, since development and deployment of new observing systems are generally very expensive and require long lead times, with strong justification for their selection desired.

Modern DAS are computationally expensive, intentionally approximate, fairly accurate, and very complex. The analyses produced by them are intended to generally be the most accurate representation of truth available. For all these reasons, a DAS can be extremely difficult to evaluate, particularly for providing insights into the mechanisms responsible for various results. Since a DAS analysis is designed to be an optimal estimate of the true atmospheric state, it is rarely possible to find better estimates to verify it. Generally, a DAS is therefore assessed by scoring subsequent weather forecasts, but this can confuse the accuracy of analyzed initial conditions with the accuracy of forecast models. Algorithms to estimate some statistics of analysis accuracy are implied by the basic Kalman filter equations (Desroziers et al. 2005), but these generally rely on unwarranted assumptions. For example, expressions for the approximate background (prior) and observation error covariances currently employed are clearly sufficiently accurate to produce analyses having whatever quality they presently do, however, determinations of analysis error covariances using the equations implied by the Kalman filter are much more sensitive to errors in those employed covariances and are therefore less reliable. Actual realizations of real analysis error generally cannot be determined.

Proposed DAS algorithms are often applied to “toy” models, such as that used by Lorenz and Emanuel (1998), to first demonstrate their utility. The models themselves provide the “truth” used for validation. The extreme simplicity of such models allows them to effectively illustrate relevant behaviors at negligible computational costs. They generally remain too simplistic, however, to provide quantitative estimates of the potential behaviors of the same assimilation algorithms applied in the context of real observations and weather.

OSSEs have been developed and validated within the context of the most modern and complete DAS and weather prediction models (Errico et al. 2017, Hoffman and Atlas 2016). As with the toy-model experiments, these also exploit a provided truth from which realizations and statistics of analysis error can be accurately determined. The DAS examined, the observations considered, the validations performed, and the weather represented, however, can be very realistic and encompassing. Since the toy model results have been considered so useful, these more realistic OSSEs should also be very useful.

The OSSE concept has been in existence for over six decades. Earliest papers from the 1950s-60s asked questions regarding observation locations, density, and accuracy, and their impacts on analysis and forecast skill. A concerted effort that used OSSEs to assess observational requirements took place as part of the Global Atmospheric Research Program (GARP), initiated in 1967. Issues such as the importance of accounting for model dependency in the OSSE were raised, and the shortcomings of the early “identical-twin” experiments, in which the forecast model and nature run were identical, were identified. Following the conclusion of GARP, OSSEs were primarily used in the 1980s to evaluate the impact of assimilating observations from expensive platforms, especially satellites, prior to those platforms being launched or deployed (e.g., Atlas et al. 1985).

A summary of early OSSEs is provided by Arnold and Dey (1986) and Atlas (1997). An uncritical review and some examples of modern OSSEs are provided in Hoffman and Atlas (2016); another review of OSSEs as applied to air quality is given by Timmermans et al (2015)). Design of a modern OSSE system is thoroughly described in Errico et al. (2017). Validation of a modern OSSE is presented in Errico et al. (2013) and Privé et al. (2013). Examples of an OSSE used to examine DAS behaviors appears in Kleist and Ide (2015a,b). In this report, some issues raised in these earlier papers will be described along with critical considerations not presented or explained elsewhere.

2. The nature run

In order to generate simulated observations and verify experimental results, some representation of meteorological fields emulating the real world is required. This simulation of nature, termed a "nature run" (NR), is generally an integration of a high-resolution climate simulation model or a long-term run of a NWP model. This provides atmospheric fields that evolve according to a formulation of atmospheric dynamics and physics. In principle, these fields can be provided at time intervals as short as the model time-step. Other sources of nature runs have been proposed, such as sequences of atmospheric analyses, but these can suffer from inadequate temporal resolution, unrealistic dynamic imbalances (or more generally, unrealistic relationships between fields), or even missing required fields, in addition to not satisfying any dynamical or physical prognostic or diagnostic equation.

a. Nature run generation

Development of a model-produced NR is generally extremely challenging. All operational DAS are applied at as high a resolution as presently practical, but for a NR, more realism, and therefore even greater resolution, is generally desired. The added realism introduces some desired portions of model and representativeness error (as explained later) as well as allows observations to be more realistically simulated. Such realism becomes more critical as the OSSE is applied to new observing systems that especially depend on features not yet modeled well; e.g., to simulate observations of precipitation, a NR requires significant realism of the mechanisms that generate rain and snow, not just the amounts produced. This consequently pushes the boundaries of current modeling capabilities into configurations beyond those already operationally or robustly tested. An extended NR testing period is thereby required.

Generation of the NR data sets is generally extremely resource intensive, in terms of both computation and, as importantly, data storage. Lengthy, high resolution NRs may take months of high-end computational processing and require many petabytes of storage space.

This also pushes the NR requirements far beyond those of current data assimilation and NWP systems.

In general, a higher spatial resolution for a NR implies the need for a higher frequency of field output. Otherwise, temporal interpolation error would swamp spatial interpolation error, reducing the total effective resolution of the NR data set. (An exception may occur for some infrequent or periodic observation types whose simulations do not require temporal interpolation.) The temporal frequency that should accompany a particular spatial resolution can be determined by estimating when magnitudes of spatial and temporal interpolation errors become similar, as described in Privé and Errico (2016). Accessing data sets having very large spatiotemporal resolution for general postprocessing and for generating simulated observations can also be very computationally expensive. For most NRs produced to date, output frequencies have been sub-optimal given their spatial resolution.

Various procedures for reducing NR data storage have been proposed by developers. This includes simulating observations during production of the NR such that they are consistent with its full spatial and temporal resolution. Archiving the observations would require less data storage space than archiving the full resolution NR fields. Alternatively, the only NR data sets archived at full resolution can be infrequent restart data sets, requiring reruns of the NR model for any, hopefully short, periods of interest. Both of these alternatives have serious drawbacks. Both past experience and future plans reveal that it is typically necessary to re-examine the NR fields at full spatial and temporal resolution to either better simulate previously considered observations or to simulate new ones, or to examine additional aspects of the NR. No clear solution to the storage dilemma is therefore apparent yet.

b. Nature Run validation

The NR must be properly validated for the intended OSSE applications. If it has been produced by an NWP or climate model, minimally this means evaluating various statistics of the model's solution (i.e., aspects of its climate) compared to corresponding statistics produced by examining real observations or analyses (as in Gelaro et al., 2015). Good agreement of all commonly-measured climate statistics is desirable but is neither a necessary nor a sufficient condition for a valid OSSE application. The actual requirements are generally different and more diverse.

Ideally, if sequences of weather, cloud cover, or other maps or depictions of NR fields are displayed alongside analogous figures produced from high-fidelity analysis of real observations, which displays are for the NR data and which are for real data would not be discernible. Note the consideration of “sequences”: realism in time is as important as realism in space. Also, a set of good NR validation metrics would quantify all the various kinds of similarity relevant. As OSSE applications change, however, the relevant set of metrics may change. Careful thought rather than simply determining a universally prescribed set of metrics is required.

As an example of this important issue, consider the simulation of radiances from a NR. If only clear sky locations are to be assimilated by the OSSE, then realistic distributions of clouds (vertically and horizontally) are desired in order to obtain realistic counts and spatial distributions of observations, particularly regarding a realistic dependence of those metrics on radiance channel. Whether the monthly and zonally mean cloud cover is realistic is effectively irrelevant for this purpose; e.g., if the cloud cover is temporally and zonally independent with correct mean-values, clear-sky radiances would always be present (or absent) at particular latitudes whereas the presence of corresponding real observations would zonally and temporally vary. Even if the means do not validate well, the temporal and zonal variations may otherwise yield realistic distributions of clear-sky locations. What is required is, therefore, consideration of what atmospheric characteristics are most relevant to validate for the specific range of NR applications envisioned.

Another important consideration is that, due to the atmosphere's chaotic dynamics, the solution of a NR model will rather quickly diverge from the sequence of real weather that occurred during the period it simulates. This implies that only statistics or frequency distributions (histograms) of the NR data set values can be validated beyond approximately 10 days (or sooner for smaller spatial scales). Also, it implies that if a particular year was chosen for the NR to simulate due to some types of events that occurred in reality, the analogous events may not be produced by the NR. As an example, the last-half of 2005 was chosen as the period to simulate using recent NRs with the hope that they would produce an extraordinary number of Atlantic tropical cyclones as occurred in reality, but that behavior was not achieved in those NRs.

The NR is run for a much longer free forecast than the shorter forecasts generally performed in operational numerical weather prediction. As the forecast integrates further than 2-4 weeks, the model may settle in to a different climatology than is seen in short or mid-range forecasts that are still significantly influenced by the initial state. It is therefore necessary to thoroughly evaluate the Nature Run dynamics and climatology, even when using a familiar numerical weather prediction model, as the behavior of the longer-term forecast may differ from previously well-characterized short and medium range forecasts.

If a new observation type or analysis algorithm is being designed for a particular class of weather events or is strongly influenced by particular atmospheric fields or specific characteristics of those fields, an OSSE naturally requires that its NR validate well with respect to those particular events, fields, or characteristics. Even if these validate well, however, it may be necessary that other NR aspects that affect the former also validate well. If observations of precipitation are being examined, for example, it may be necessary that not only the NR precipitation field look realistic in both time and space but that the relationship between that field and others that generate the precipitation also appear realistic.

The qualifications offered in the previous paragraph are purposefully stated because the degree of validation required depends strongly on what metrics are to be applied and the nature of the influences on the observations. If a particular field acts primarily to determine DAS data selection and quality control (QC) decisions, such bi-modal responses (select or omit an observation) can generally be easily mimicked when simulating observations for an OSSE regardless of how well those determining fields are represented in the NR. This of course may be true only up to some degree, and generally it would be most useful were the NR realistic in all aspects. The point is, that the requirements are not as straightforward as they may otherwise seem. Careful thought and understanding of the key issues peculiar to the particular OSSE are required.

A NR production may be constrained by real data to some degree. As an example, the G5NR (Putnam et al. 2014) used prescribed sea surface temperature and sea ice fractional coverage fields as well as aerosol sources derived from observations for 2005-2007. It also used initial conditions for soil wetness derived from such data. If the OSSE DAS uses some of the same data, even if applied at a different resolution, it can misjudge what could otherwise be sources of error. It is critically important, therefore, to know which constraints have been applied to the OSSE and to consider what their implied effects may be in a particular OSSE context.

3. Simulated observations

Simulated (or synthetic) observations are created to provide a set of data for ingestion into a DAS that mimics observations that would be made if the Nature Run were the real world. This is generally accomplished by applying some observation operators to interpolations of the Nature Run fields to spatiotemporal locations where an actual instrument might observe. For some data types, this may be a simple and straightforward process, while for others it may be much more difficult, as described in what follows.

An operational DAS does not use all observations provided to it. Data for some observation classes will be thinned to remove redundancies, filter correlated observation errors, or to reduce computational or data storage requirements. Sometimes this thinning depends on values of the observations themselves, and not just their locations in time and space; e.g., it may preferentially retain radiance values that appear less cloud-affected. All data are also subjected to quality control to remove likely gross observation errors that do not otherwise fit the Gaussian error model specifically considered by most DAS (Lorenc and Hammon, 1988). Nonetheless, some data actually assimilated will contain gross errors that were undetected, although QC schemes will generally constrain such retained errors to be small. Other observational errors that are not considered gross will generally be retained, although perhaps partially filtered, by the DAS.

There are therefore several critical properties of the simulated observations whose realism can strongly affect the DAS results. These include numbers of observations, spatial and temporal distributions of observations, relationships between observations and the fields to be analyzed, and characteristics of instrument and representativeness errors, including biases, variances, correlations, and their relationships to local synoptic conditions. Data thinning and QC affect these properties. Once those procedures are applied by the DAS, however, details of the excluded observations (i.e., regarding their temporal and spatial distributions and the distributions of their gross errors) are unimportant. At this point of their application, realism of the observations specifically refers only to those accepted by the thinning and QC since only their values are applied to the DAS's solution (data-fitting) algorithm.

For the above reason, it is therefore extremely helpful to distinguish between the characteristics of the observations that particularly affect their thinning and QC in contrast to those that affect results of the DAS solution algorithm. Some details of the former characteristics may be unimportant. The sizes of gross errors or the pre-QC numbers of provided observations do not matter for those rejected. It may be important, however, that the retained simulated observations are in cloudless regions (when assimilating clear-sky radiances) or in cloudy regions (when simulating cloud-tracked winds) since such different regions are associated with different kinds of weather and forecast error characteristics. It is therefore sufficient to simulate the observations such that realistic distributions and counts occur, but not necessarily that the rejected observations have characteristics similar to real rejected ones.

As an example of the previous issue, consider atmospheric motion vectors (AMVs) and clear-sky radiances. The NR may not have very realistic cloud distributions. Thus actual cloud-tracking algorithms or cloud scattering radiative transfer algorithms applied to the NR fields will not produce realistic distributions of simulated AMV or cloud-affected radiance observations. If the first goal of the OSSE is to simulate cloud-associated observations where the NR has clouds, with realistic counts and spatial separations, then many details that would be provided by these algorithms are unnecessary; e.g., whether the effect of cloud scattering on an IR radiance is to reduce the brightness temperature by 10 C instead of 20 C, is irrelevant since either value will likely be eliminated by thinning or QC. What may be relevant, however, are the numbers of cloud-affected observations that remain undetected and the characteristics of any remaining errors. Such numbers may be small if the QC is rigorous. So, when simulating observations adequate understanding of the nature of gross errors and of QC and thinning procedures is required, although the gross errors themselves need not be simulated well.

Realistic simulation of a specific observing system does not require a uniform degree of realism across all facets of the modeled atmosphere. For example, it is not so important that clouds be simulated well if an observation is unaffected by them. In contrast, observations that either are intended to measure clouds or are strongly affected by them will require a high degree of realism in the simulation of the clouds. As more realism of the simulated observations is required, the NR may need more realism also.

Observation impacts may be grossly misrepresented if the simulated observation is mischaracterized. One example of this is when radiance observations are generated from a NR as though they are radiosonde profiles and then assimilated by the DAS as radiance retrievals but without considering how real retrievals and their resulting errors are impacted by the algorithms that construct them (Joiner and Da Silva 1998). Another example occurs when GPS radio occultation observations are generated and assimilated as refractivity measurements but without considering the processing errors inherent in such derived quantities. A third example regards dropsondes within NR tropical cyclones, where the sonde trajectories should be simulated using the high-speed winds present. Such mischaracterizations often occur when there is a rush to obtain OSSE results without appropriate consideration of the nature of the observations in their various forms (particularly their observational errors) and of the implications of simplifications introduced primarily for computational ease.

4. Simulation of observation errors

The values of most metrics used to assess real atmospheric data assimilation or forecast systems are determined by the various types of errors inherent in those systems and how the modeling and data assimilation modifies them. These types include observation instrument errors, observation representativeness errors, forecast model formulation errors, and errors introduced by the data assimilation algorithm itself. The last includes phase errors introduced by an incremental analysis update scheme or bias correction errors due to a wrong partitioning of bias between observation and background values. Of course, the magnitudes, shapes, and locations of all these errors change as the model dynamics and physics acts on them and the data assimilation system attempts to filter them. Validation of an OSSE therefore depends on how well all of these errors and the processes affecting them are simulated. In particular, how well the observations are simulated actually concerns how well their instrument and representativeness errors are simulated, especially regarding those characteristics impacting analysis and forecast quality. Although these errors have often been neglected in OSSE studies, the theory underlying data assimilation clearly states that errors in the DAS products, and thus the reliability of those products, particularly depend on errors in the DAS observations and model.

Observational errors may be separated into two sources. One is the error in the measurement itself. This is generally termed the "instrument error" although it can include error produced when processing the data before its dissemination; i.e., a component of error not actually induced by a physical device. The other is termed "representativeness error" or "forward model error". This arises due to errors in the way the observation is quantitatively related to the fields to be analyzed within the DAS. In a sense therefore, these latter errors are induced by the particular DAS algorithms employed, although within the DAS theory (and mathematics) they are properly combined with the instrument errors (Tarantola 1987, Lorenc and Hammon 1988). For many types of observations, magnitudes of representativeness error surpass those of instrument error. Thus claims such as "measurements with the proposed instrument are error free" may be irrelevant, especially if the forward model required to assimilate them is quite complicated or indirect.

One example of representativeness error is that induced by inaccuracies of the fast radiative transfer schemes generally used to relate observed radiances to fields of temperature, humidity, ozone, etc. Such errors can be greater than instrument measurement errors; e.g., when surface emissivity is incorrectly specified or radiative scattering is ignored. Another example of representativeness error is that due to spatial or temporal interpolation between gridded fields. Sometimes it is only this type of representativeness error so termed, with the earlier mentioned type specifically called either forward model error or observation operator error. Both types of representativeness error, however, enter the DAS problem mathematically similarly, and the interpolation error can be simply considered as one caused by another (i.e., interpolation) model or operator.

Actual instrument error is obviously not present in an OSSE since there is no real instrument present. Such error therefore needs to be simulated and added to the observation. A forward model is present, however, as it is applied to the NR to produce the observations. If this forward model is not identical to the one in the DAS or is applied to a different grid structure or to different types of fields, then some representativeness error is implicitly produced. Any such differences are interpreted by the DAS in the OSSE context as representativeness error. In general, however, the statistics of these differences will not be the same as those between the DAS forward model and the real physics, spatial, or temporal relationships the forward model represents. The typical sizes (e.g. variances) of the real differences are likely larger than the differences between the NR and DAS applications. It is only such discrepancies that need to be explicitly simulated, along with the instrument error, and added to the simulated observations to render their errors realistic.

In a critical sense, the nature of representativeness error in OSSE and real contexts is very different. This is described thoroughly in Errico et al. (2017). Essentially, the distinction is this: as a more realistic observation forward model or better interpolation scheme is applied in a DAS ingesting real observations, the representativeness error will generally decrease (this is the meaning of "better"). For a fixed DAS ingesting observations simulated from a NR, however, as these observations are simulated more realistically (and therefore more differently) compared to what the corresponding observation operators in the DAS would produce, the representativeness error will generally increase. It should be understood, therefore, that when greater realism is introduced in the OSSE simulation or NR compared with what is used in the DAS, the information in the measurement that can be utilized by the DAS is not increased, but the observational error is!

Although simulating observations for an OSSE is critically about simulating their errors, not all error characteristics are equally important. Some types of errors are effectively removed by a DAS, so they have negligible effect on either analysis or forecast error statistics. One example is any specific observational bias that is accurately determined by a bias-removal algorithm in the DAS. Another is uncorrelated random error that the DAS is designed to remove. Although inclusion of such errors may be required to validate some DAS statistics obtained from an OSSE, such as innovation biases (before DAS bias correction) or innovation standard deviations, they will have less impact on either analysis or forecast error validation metrics as the DAS effectively removes them.

Most current DAS are designed to especially filter random uncorrelated errors from observations. This they do very effectively as long as the spatial densities of relatively independent observations remain high. Most correlated errors, however, will be retained by the system. The filtered uncorrelated errors therefore have little effect on the analysis since they are effectively removed by the DAS, but the unfiltered correlated errors effectively remain to impact the analysis. For an OSSE to realistically mimic the effects of retained errors, it is therefore imperative that the simulated observation errors particularly include any significant correlations found in real observation errors.

When it can be reasonably assumed that the background error variances and correlations in the OSSE and real contexts are similar, then any large differences between corresponding innovation correlations can be attributed to unrealistic statistics for the OSSE simulated observation errors. Once the observation error variances are tuned so that the innovation variances sufficiently match, observation error correlations can be introduced so that the innovation correlations also match (Errico et al 2017). If the background error statistics poorly match, however, compensating for them by instead adjusting the observation error variances or correlations may be unproductive for simultaneously matching other aspects of the DAS behavior, since the cause of the discrepancy is thereby ascribed to the wrong source. For example, if the variance of background error is lower in the OSSE than with real data due to insufficient model error, the variance of observation innovations will also tend to be lower in the OSSE. (The innovation variance is the sum of the true observation error variance and the background error variance in observation space, if the two types of error are uncorrelated.) In order to match the variance of observation innovation, the variance of the observation error would need to be artificially inflated to a higher level than for real observations in order to compensate for the smaller variance of background error.

If adequate observation errors are not added, impacts of those observations on improving analysis or forecast skills will most likely be overestimated. The discrepancy may be small if other errors (such as model error) dominate the overall error, but then the observation impact itself may be small. Examples of these various results appear in Privé and Errico (2013b). Similarly, if the observations are spatially or temporally dense and the DAS is designed to only remove uncorrelated observation errors, if the real observation errors are highly correlated but the simulated errors are not, the observation impacts will also likely be overestimated. Again, the discrepancy may be small if other sources of error dominate. In general, however, the potential effects of error correlations should always be considered, even if they cannot be properly simulated.

5. Forecast and Forward Models

The DAS and its observation operators are used to ingest the simulated observations and perform the OSSE experiments. Ideally, the DAS observation operators should be different than those used to generate the synthetic observations so that realistic representativeness error is introduced. Also the DAS forecast model should be substantially different from the model used to generate the NR so that realistic forecast model error is introduced. Here, the latter error specifically means those errors due to the imperfect formulation of the forecast model used to advance a sequential analysis, not the total error of a forecast that depends on the combined effects of errors due to model formulation, initial condition, and boundary condition errors. In the real data assimilation context, the forecast model error is determined by differences between the approximate physics and dynamics specified in the computational forecast model versus the complete physics and dynamics of the real world. The OSSE equivalent of this error is the difference between the model formulations (including grid structures) used for the NR and the DA system. In practice, however, differences between corresponding observation operators or forecast models are generally smaller than those between either and the real world. The simulated errors such differences implicitly introduce are therefore generally under-estimates of real ones.

One type of OSSE that is often seen in the literature, particularly in older studies, is the “identical twin” or perfect model OSSE, in which the same model is used to generate the NR and to provide the DAS and OSSE forecasts. Substantial caveats arise, however, when using an identical twin framework to evaluate present or future observation systems or DAS algorithms. The resulting analysis is expected to have much lower error than in the real world, substantially affecting the observation impact on subsequent forecasts. While some efforts may be undertaken to artificially degrade the analysis state to include more error, this is fraught with danger as the magnitude and distribution of analysis error in the real world is not well understood. In particular, it is expected to be highly correlated spatially and temporally as well as dependent upon the atmospheric state. While an identical twin OSSE nonetheless may be used successfully for certain theoretical experiments or carefully designed proof of concept trials, the perfect model setup should be avoided for any reliable quantitative evaluation of new observing systems.

Insufficient model error is a difficult problem that plagues all OSSEs. Even if the variance of true model error is mimicked in the OSSE, equally critical aspects such as error correlations or dynamical balances, may not be. Inadequate model error can affect not just forecast skill and observation impact measures, but nearly every aspect of the OSSE performance, including the calibration process in which the OSSE statistics are tuned to replicate the real world. In particular, without sufficiently realistic, simulated model error, the actual background error statistics encountered in the real versus OSSE DAS contexts will likely differ. Realistic observation error covariances can then not be simply estimated as residuals by assuming that the measurable background error variances in the OSSE context are nearly identical to the poorly known background error variances in the real context, as currently considered for some OSSEs (Errico et al. 2017). Also, the static component of the background error covariances prescribed for a real DAS (i.e., its specified "**B** matrix") may be less appropriate for an OSSE applied using the same observing system. With the difficulty of tuning static background error covariance models, such discrepancies should be avoided if possible; otherwise the OSSE DAS should be considered more sub-optimal than the real DAS with appropriate disclaimers regarding interpretations of its results.

For a numerical weather prediction system that is stably cycling, the growth or decay of errors (due to both model error and dynamical modification of initial condition error) during a cycle period is statistically balanced by corrections or errors provided by the ingested observational data. In the real world, it is difficult to distinguish between the results of model error and initial condition error. If the OSSE simulated observation errors are tuned so that the net error growth in the OSSE and real world match, it is likely that the observation errors will be over-inflated to compensate for the reduced model error. Even if this can be done to successfully match results for analysis cycling periods, it will not mitigate effects of insufficient model error being applied during the subsequent OSSE forecasts.

One method of dealing with the issue of insufficient model error is to deliberately increase the model error by altering the physical or dynamical processes of the DAS model compared with those of the given NR model. For example, the convection scheme or physics parameterizations can be changed so that there is greater difference in the behavior of the two models. However, this method can be criticized as deviating from the goals of the OSSE, which are to have a NR that is as close as possible to reality, and the experiment model as close as possible to an operational NWP model. Also, without extensive retuning of the modified model, new unwanted deficiencies should be expected. In practice, only a small portion of the originally deficient model error will likely be corrected by minor model changes. Careful testing of any intentional adjustments to the NR or forecast model therefore should be made before using the altered model in an OSSE.

Of especial interest is the impact of insufficient model error on measures of observation impact in the OSSE. There is a particularly complicated interaction between observation impact and model error that can make it difficult to make a general statement about the expected effects. On the one hand, lack of model error leads to less "work" for observations to perform in correcting the initial backgrounds. This can commonly be seen in adjoint estimates of observation impact, which may be up to 50% smaller in a fraternal-twin OSSE compared to an estimate for real observations. On the other hand, the impact of observations may be retained further into a forecast integration when there is less model error that will tend to negate that impact. So while the overall short-term impacts of observations are likely to be reduced in an OSSE, some impact metrics may be greater in an OSSE than would be measured for corresponding real observations. Although estimating such impacts with an OSSE is often its primary purpose, those estimates therefore often have some ambiguity.

6. Validation

Since an OSSE is fundamentally a simulation of the application of a DAS to real observations, it is critical that the OSSE is able to adequately mimic whatever properties of a real-data DAS are relevant for a given application. The goal is always to use the OSSE to inform us about reality. Because every aspect of the OSSE is simulated, there is no way of knowing how well the simulation represents reality without some verification process. The characteristics or metrics to be investigated vary with the application, as with the NR validation. Adequate validation is required, regardless of attempts to apply unrealistic time constraints to the OSSE development. Without this, claims of the OSSEs scientific basis should not be offered.

The validation may reveal problems with the OSSE setup that can be resolved with minor changes, or the validation may expose more deep-seated issues that cannot reasonably be mitigated. In the latter case, the validation will show how far the results of the OSSE can be trusted, and what caveats should be attached to any problematic aspects of the study.

Validation of the OSSE requires examination of statistics. Specifically, these are statistics that can be measured in both OSSE and real data contexts. As an example, statistics of observation innovations that are routinely monitored in a DAS, can be directly measured in both contexts. In contrast, although analysis error or Kalman gain can be explicitly measured in the OSSE, they can only be unreliably estimated in the real data context. The latter metrics, whose qualitative reasonableness should be assessed, are therefore not among those that can be used for quantitative validation.

A great variety of statistics should be examined. These include counts, means, variances, and correlations of observation innovations as functions of geographic locations and levels or channels. Also included should be means and variances of analysis increments and of forecast errors evaluated with verification using the fields of a subsequently-produced analysis as truth rather than the Nature Run. Such "self-analysis" statistics can be computed in both real and OSSE contexts. Also desirable is computation of observation impacts, either by performing observing system experiments (OSEs) or by especially using forecast model and DAS adjoints (See Gelaro and Zhu 2009 for a comparison of the interpretations using the two methods).

In the OSSE, analysis, background, and forecast error statistics should also be computed with respect to the truth provided by the NR (e.g., Errico et al. 2013 and Privé et al. 2013a). This is not as trivial as it appears because generally the OSSE fields are on a different grid, with different resolution, than the NR fields. The truth and OSSE fields must first be put on a common grid or the precise relationship between them defined. This requires consideration of questions such as "Do provided grid point values denote point values or grid box mean values?" and "Is it desirable to include in the quantified analysis error that portion due to the inability of the analysis to represent the full resolution of the NR, or just the portion that it can represent?" Answers to such questions as these determine how the verifying truth should be computed and interpreted. The smaller the analysis or forecast errors, the greater the relative portion of error that is affected by such details.

7. Experiments and expectations for results

Without an actual data set that represents an error-free truth, realizations of real analysis or background error cannot be determined. Without such truth, neither can such errors be corrected in a definite sense. The data assimilation problem is therefore fundamentally a statistical problem. It is only in a statistical sense that errors can be characterized or measured, and any metrics of general performance must be statistical measures.

One corollary of this statistical nature is that error sampling issues must be considered. A singular or small set of case studies, although potentially illustrative, will have only limited utility. In particular, to discern the small improvement of analysis or forecast accuracy expected for any reasonable augmentation of the currently observed atmosphere with any statistical significance will generally require very long periods of examination.

There are some important limitations constraining what can be learned from an OSSE. First, DA results depend on how observations are used, not simply how good they might potentially be. Generally, the more revolutionary a new data source may be, the less a current OSSE may be able to demonstrate its utility. As an example, consider proposed observations that are affected by precipitation such that they can, in principle, aid in analyzing or predicting precipitation. A DAS that does not adequately address many of the critical issues that render precipitation analysis so difficult (Errico et al. 2007), however, will likely fail at its use of even good information. The resulting poor impacts of those observations therefore would indicate our current inability to utilize those observations rather than their potentially greater utility when the future observing system is finally deployed. By that time in the future, the critical assimilation issues may have been addressed.

Second, there are currently hundreds of millions of observations of the earth’s troposphere and lower stratosphere daily, mostly from globally-viewing satellites. Modern DAS utilize tens of millions of these observations per day after choosing what appears to be only the best after thinning and QC. Augmentation of this existing network with a new observation set will likely only make modest improvements to analysis and forecast accuracy. The fact that forecast model error remains significant further limits the potential impacts of new observations because only a portion of the total error is due to observation error.

Third, given the chaotic nature of atmospheric dynamics, using a DAS even with only marginal observations is much better than having no observations. Without them, a cycling global DAS behaves the same as a long-term global forecast: it is only constrained by the climatology of its model, and its sequences of weather patterns will tend to diverge from those either seen in nature or produced by a NR using a different model. So, although a new observation type may be sufficient to profoundly constrain a cycling DAS to some degree, its individual impact within the context of a realistic suite of observations may instead be minimal. While an investigation of an observation type in an OSSE with few other observations may be informative, the results should not be misinterpreted as revealing its likely impact in a DAS utilizing a complete set of observations. An OSSE with such a complete set should instead be performed to estimate the latter impact.

OSSE results may be desired for new instruments that may not be made operational for a decade or more. Greater caution is needed when interpreting the results of OSSEs for such future missions, as substantial changes to the global observing network, the DAS, and the forecast model are likely to occur during the intervening time. While some accommodation of expected changes to the observing network can be made in the OSSE, improvements in the modeling system are much more difficult to anticipate. Since a DAS determines the weighting given to each data type, changes to the data assimilation algorithm can change those weights and thus alter the impact of each data type. Also, improvements to the forward model can make a particular type of data more or less useful. For example, a cloud resolving model may use observational data of humidity and cloud characteristics differently and more effectively than a model that does not resolve cloud structures.

For these reasons, hopeful providers of a proposed observing system should be informed that a legitimate OSSE will likely not deliver the extraordinary impacts that would so greatly assist in promoting their system. As a corollary, the use of more poorly designed and conducted OSSEs may create more encouraging results, yielding a marketing advantage to competing data providers who employ them. Some kind of oversight is therefore urgently required to mitigate the potential for this otherwise likely confusion.

The stakeholders who are relying on the results of the OSSE to inform the decision-making process should be advised about the limitations and qualifications of OSSEs in general and for the specific OSSE being performed in particular. Ensuring that decision makers have reasonable expectations regarding the time required to perform a robust OSSE and the scope of applicable results will help support the credibility and usefulness of the OSSE.

8. **Achieving Robust Results**

No model is a perfect replica of the real world, and as such there is no such thing as a perfect OSSE. In reality, meeting all of the best practices for developing and performing an OSSE is not possible due to limited resources and time. This does not mean that OSSE efforts should be abandoned. Instead, extra care should be taken to estimate and understand the likely impacts of deficiencies in the implementation of the OSSE on the experimental results. Seemingly large shortcomings in the OSSE framework may have only minimal influence on the experiment, while more subtle limitations may substantially corrupt the results.

Verification of the OSSE performance is critical for proving the trustworthiness of the OSSE, not just for the researcher but also for the stakeholder. A minimum guideline for the verification process is to first carefully consider the most critical aspects of the desired experiment before embarking on the OSSE. The most important metrics for all of these critical aspects should also be determined. These diagnostics then form the basis of the verification procedures that will be used to validate the OSSE. Each application of an OSSE framework therefore requires some unique considerations, beyond following universal recipes.

a. Working with Case Studies

Case studies can be particularly challenging in an OSSE. In the real world, past events may be selected for investigation based on knowledge of the behavior and impact of the event as well as the forecast skill for the event. For an OSSE, the availability of events of interest is generally unknown until after the Nature Run is generated, and there may be only a very limited number of events to choose from. The forecast skill for each of these events is also unknown until after all of the development effort for the OSSE has been completed and experiment forecasts are performed. An event that appears promising may have highly skilled forecasts in the control experiment, leaving little room for improvement due to new observations.

The NR may not realistically represent all aspects of a particular type of event. For example, the distribution or magnitude of precipitation in an extreme event may be poorly modeled. Tropical cyclone intensity and structures can be difficult to model, as can orographic effects and other highly-localized phenomena. Although the NR may have realistic representations of some aspects of a type of event or phenomenon, other aspects may be inadequately rendered. Before experiments are initiated, the NR event(s) should be carefully analyzed to ensure that the behavior of interest is sufficiently realistic. This includes comparison of the desired experimental metrics against similar real world metrics, if possible.

Case studies can be hindered by deficiencies and shortcuts that may otherwise be acceptable for computing global metrics or long-term statistics. For example, if the synthetic observations of radiances or atmospheric wind vectors are not affected by NR cloud fields, the simulated data may be located in regions where data should not be present, and absent in regions that should contain observations. This could substantially change the analysis and forecast impact of experimental new data, leading to either underestimation or overestimation of impacts. If resources are not available to correct these deficiencies in the synthetic observations, at a minimum the simulated observations should be overlain with the NR fields at the observation times to identify any discrepancies. Any particularly troublesome simulated observations could be manually removed or altered, or at the very least the interpretation of the experimental results should be strongly tempered.

Sometimes a short NR is generated specifically for a particular real world event, with the run initialized from a time shortly before the event begins in order to ensure the existence of a similar event in the NR. This approach has several advantages. The behavior of the NR can be directly compared to the real world event for validation; likewise, simulated observations and their assimilation metrics can be compared with corresponding real observations. Even in this case, the new observations to be tested in the OSSE have no real world equivalent and sample the NR in ways that cannot be directly verified. Another major challenge is that if the NR model is capable of generating the very realistic and accurate forecast desired, there may be little room for improvement in the experimental forecasts.

Case studies can be useful for illustrating either problems or promises of new data assimilation or observing systems. Since all data assimilation algorithms are necessarily derived from fundamentally statistical considerations, however, single, short case studies do not generally provide sufficiently robust statistical samples from which to draw confident conclusions. This is especially true when the OSSE application has been effectively “tuned” to yield desired results.

b. Regional OSSEs

Sometimes it is desirable to perform an OSSE using a forecast model and NR that are at such high resolution as to be untenable for a global model. In this case, a regional OSSE is often performed. While it may appear that development and application of a regional OSSE would be less work than a global one, the opposite is actually true. In order to perform a regional OSSE that covers more than a day or two, both a global NR and OSSE and an embedded regional NR and OSSE should be generated. This is principally because the lateral boundaries very quickly affect the interior of the regional model solution and the use of otherwise perfect boundary conditions unrealistically constrains the forecast too completely (Errico and Baumhefner 1987; Denis et al. 2002).

A global NR can provide lateral boundary conditions for the regional model NR and a global OSSE can provide realistic errors of those boundaries. An alternative that is often seen in the literature is to use a series of global model analyses instead of a free-running model NR. The simplicity of this is enticing, especially when a new observing system being examined is to be deployed entirely within the regional model domain, far from its lateral boundaries. Global analyses are generally available only a few times per day and therefore may be temporally inadequate to provide good lateral boundary conditions. They also suffer from the other problems previously noted when attempting to use them as NRs. Additionally, if the new observing system is also expected to be deployed outside the regional domain, its potential to improve the regional lateral boundaries would be missed unless a global OSSE is simultaneously conducted to affect them.

A regional NR generally at higher resolution than the global NR, should produce features that correspond well to those in the global NR so that gross mismatches at the lateral boundaries do not occur. For example, a front or tropical cyclone should be colocated in both the global and regional NR. For regional NRs that encompass a large region or are run for a significant length of time, this may be challenging. Nolan et al. (2013) give a good description of the process of creating an embedded regional NR within a global NR.

9. **Additional caveats**

Several suggestions to expedite the OSSE development and application process have been offered. Most involve some combination of circumventing adequate validation, employing inadequate nature runs, ignoring representativeness errors, ignoring important characteristics of the observations (e.g., treating radiance retrievals as radiosondes, even when the number of channels is very small, without considering the resulting profound vertical correlations of errors) or ignoring the fundamentally statistical nature of the DAS problem. Reasons are generally offered to the effect that some administrator or client demands quick results, regardless of the request's reasonableness. If an OSSE is to be considered truly useful and a proposed observing system is expected to last a decade with a total cost of tens of millions of dollars, however, it follows that the OSSE used to promote that system should be reliable. Also, conducting a sub-standard OSSE but offering many caveats should be considered equally unacceptable. Of course all OSSEs will be imperfect and caveats will always be required, but they should not be simply offered because the experimenter is not sufficiently knowledgeable or only has access to an inadequate system. Some minimal standards should be met.

Offers have been made to provide the OSSE community with a "black box" capability, whereby "a graduate student" can select among some choices on a graphical interface screen and conduct an OSSE. This could provide many combinations of options and greatly facilitate OSSE applications since little development time or experience would be required. For the same reasons, however, the potential for misuse of OSSEs would also be greater. Given the complexity of modern DAS, the generally indirect relationship between all observational data and what we are attempting to analyze, the additional complexity of the OSSE and NR application, and the large sums of money to be invested in designing, manufacturing, and deploying new observing systems, the potential for misuse of "black box" OSSEs is significant and must be minimized. The WMO community would be better served if there are a few adequately supported groups with appropriately trained personnel to conduct such work. As has been extraordinarily well demonstrated within the global DAS/NWP community, the several such groups should be making continuous comparisons with each other, freely exchanging criticisms, algorithms, and recommendations.

Data assimilation is a truly interdisciplinary problem. It involves statistics, linear algebra, computation, instrument design, modeling, physics, and dynamics. Performing an OSSE requires at least a rudimentary knowledge of all these aspects. It requires a greater understanding of those aspects that are expected to determine results of the specific experiments. For this reason, OSSE researchers should have adequate training in general data assimilation and also have ready access to relevant experts.

Finally, it is essential that OSSE conductors avoid conflicts of interest. This clearly includes having financial stakes in the promotion of a new observing system. More subtle but as damaging are experimenters acting as obligated to "doctor" their experiments in order to satisfy the desires of the observation designers paying for the experiments. Such concerns are not theoretical. They have occurred in the recent past. Instead, OSSEs must be funded and administered in ways that would render these concerns unlikely.

10. Benefits of OSSEs

The development and validation of an OSSE framework is an undertaking that requires substantial resources and effort. However, once this OSSE framework has been created, there are many different ways in which the OSSE may be employed that require relatively minor effort. A wide range of different instrument types may be investigated for analysis and forecast impact in a well-developed OSSE system. The same OSSE may be used to experiment with various aspects of behavior of the DAS and forecast system.

For new observation types that may be deployed soon, much of the additional OSSE development required to explore them in a pre-deployment OSSE context will expedite their utilization after real observations become available. This includes the ability to read, select, quality control, utilize, and validate the coming data. The OSSE framework can also be used to help adjust the DAS to maximize the potential impact of the new data and to anticipate or eliminate some potential problems. Thus the OSSE investment should not be considered as entirely ancillary to the utilization of new observing systems.

The use of "toy" assimilation systems and models to examine DAS algorithms has been ubiquitous (e.g., see Anderson 2016). These are computationally exceedingly inexpensive and provide their own "truth" for precisely measuring impacts. Even developers working with operational systems have toy models useful for understanding the otherwise always complicated and sometimes unintuitive behaviors encountered in DAS applications. Although an OSSE system designed and validated for operational forecast models, observations, and assimilation systems is not as negligibly expensive, it also provides a truth to provide precise realizations and statistics of analysis errors. Unlike the toy systems, however, their simulations can be almost entirely realistic and provide a context within which almost all aspects of the assimilation problem can be investigated. OSSEs can therefore provide realistic estimates and not just illustrations of real DAS results. They therefore should prove at least as valuable as investigations using toy systems.

11. Summary and Recommendations

The presentation in this report has focused on OSSEs for NWP due to the limited expertise of its authors. The procedures, recommendations, and warnings offered, however, should apply to OSSEs for most other applications, especially those for the earth sciences. The basic principles of DA are indeed quite general. So also are the principles of the scientific method.

Development and applications of OSSEs should be strongly encouraged. This does not mean that they should be applied to all proposed DAS algorithms or observing systems. OSSEs are likely not sufficiently mature for all investigations desired. The appropriate purposes are sufficiently many and critically important, however, that the modest resources required should be provided to a few groups both willing and able to develop, validate and apply such experiments.

Results provided by OSSEs can profoundly impact the entire community by shaping the future earth observing system. Also, the developmental problems encountered before OSSEs can be applied are quite complex and interdisciplinary. For these critical reasons, OSSEs should be conducted by researchers having access to the required software and computational resources who are willing to exercise the significant diligence, educational investment, judgment, etc., required. On the other hand, it is vital that multiple efforts exist. Only then can extensive development efforts be distributed and shared, and only cooperative competition can provide opportunities for cross validation and constructive criticism. The products produced in such an environment will be more robust and timely than otherwise possible.

As with all truly scientific endeavors, ample public validation and constructive criticism should be encouraged. This does not simply mean that opportunities for showcasing results should be provided at popular meetings, as already occurs. Instead, venues and formats must provide, and actually encourage, opportunities for ample questions, discussion, and even debate. Only then will confidence in OSSE results and conclusions be scientifically warranted.

12. References

Anderson, J.L., 2016: Reducing correlation sampling error in ensemble Kalman filter data assimilation. *Mon. Wea. Rev.*, **144**, 913-925. doi: 10.1175/MWR-D-15-0052.1.

Arnold, C. P. and C.H. Dey, 1986: Observing-Systems Simulation Experiments: past, present, and future. *Bull. Amer. Met. Soc.*, **67**, 687-695.

Atlas, R., E. Kalnay, and M. Halem, 1985. Impact of satellite temperature sounding and wind data on numerical weather prediction. *Opt. Eng.*, **24**(2), 242-341.

Atlas, R, 1997. Atmospheric observations and experiments to assess their usefulness in data assimilation. *J. Meteor. Soc. Japan*, **75**, 111–130.

Denis, B., R. Laprise, D. Caya, J. Côté, 2002. Downscaling ability of one-way nested regional climate models: the big-brother experiment. *Climate Dynamics*, **18**, 627-646. doi 10.1007/s00382-001-0201-0.

Desroziers, G., L. Berre, B. Chapnik, and P. Poli, 2005. Diagnosis of observation, background, and analysis-error statistics in observation space. *Quart. Journ. Royal Met. Soc.*, **131**, 3385–3396. doi:10.1256/qj.05.108.

Errico, R.M., and D.P. Baumhefner, 1987: Predictability experiments using a high-resolution limited-area model. *Mon. Wea. Rev.*, **115**, 488-504.

Errico, R.M., P. Bauer, j.-F. Mahfouf, 2007: Issues regarding the assimilation of cloud and precipitation data. *J. Atmos. Sci.*, **64**, 3737-3741.

Errico, R. M., R. Yang, N. Privé, K.-S. Tai, R. Todling, M. Sienkiewicz, and J. Guo, 2013. Development and validation of observing-system simulation experiments at NASA's Global Modeling and Assimilation Office. *Q. J. Roy. Meteor. Soc*, **139**, 1162-1178. doi: 10.1002/qj2027.

Errico, R.M., N.C. Privé, D. Carvalho, M. Sienkiewicz, A. El Akkraoui, J. Guo, R. Todling, W. McCarty, W.M. Putman, A. da Silva, R. Gelaro, I. Moradi, 2017. **Description of the GMAO OSSE for Weather Analysis Software Package: Version 3**. *NASA Technical Report Series on Global Modeling and Data Assimilation, NASA/TM-2016-104606,* Vol. **48**, 156 pp.

Gelaro, R. and Y. Zhu, 2009. Examination of observation impacts derived from observing system experiments (OSEs) and adjoint models. *Tellus,* **61A**, 179-193.

Gelaro, R., W. M. Putman, S. Pawson, C. Draper, A. Molod, P. M. Norris, L. Ott, N. Privé, O. Reale, D. Achuthavarier, M. Bosilovich, V. Buchard, W. Chao, L. Coy, R. Cullather, A. da Silva, A. Darmenov, R. M. Errico, M. Fuentes, M.-J. Kim, R. Koster, W. McCarty, J. Nattala, G. Partyka, S. Schubert, G. Vernieres, Y. Vikhliaev, and K. Wargan, 2015. **Evaluation of the 7-km GEOS-5 Nature Run**. *NASA/TM–2014-104606*, Vol. **36**.

Hoffman, R.N. and R. Atlas, 2016. Future Observing System Simulation Experiments. *Bull. Amer. Met. Soc.*, **97**, 1601-1616. doi: 10.1175/BAMS-D-15-00200.1

Joiner, J. and A.M. Da Silva, 1998. Efficient methods to assimilate remotely sensed data based on information content. *Quart. J. Roy. Met. Soc.,* **124**, 1669-1694.

Kleist, D.T. and K. Ide, 2015a. An OSSE-based evaluation of hybrid variational-ensemble data assimilation for the NCEP GFS. Part 1: system description and 3D-hybrid results. *Mon. Wea. Rev.*, **143**, 433-451.

Kleist, D.T. and K. Ide, 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.

Lorenc, A.C. and Hammon, O, 1988: Objective quality control of observations using Bayesian methods. Theory and a practical implementation. *Q. J. R. Meteorol. Soc.*, **114**, 515-543.

Lorenz, E.N. and K.A. Emanuel, 1998. Optimal sites for supplementary weather observations: simulation with a small model. *J. Atmos. Sci.*, **55**, 299-414.

Nolan, D. S., R. Atlas, K. T. Bhatia, and L. R. Bucci,2013. Development and validation of a hurricane nature run using the joint OSSE nature run and the WRF model, J. Adv. Model. Earth Syst., 5, 382–405, doi:10.1002/jame.20031.

Privé, N., R. M. Errico, and K.-S. Tai, 2013a. Validation of forecast skill of the Global Modeling and Assimilation Office observing system simulation experiment. *Q. J. Roy. Meteor. Soc.*, **139**, 1354-1363. doi: 10.1002/qj.2029.

Privé, N., and R. M. Errico, 2013b. The role of model and initial condition error in numerical weather forecasting investigated with an observing system simulation experiment. *Tellus-A*, 65, 21740. doi: 10.3402/tellusa.v65i0.21740.

Privé, N. C., R. M. Errico, and K.-S. Tai, 2013. The influence of observation errors on analysis error and forecast skill investigated with an observing system simulation experiment. *J. Geophys. Res. - Atmos*, 118, 5332-5346. doi: 10.1002/jgrd.50452.

Privé, N. C., and R. M. Errico, 2016. Temporal and spatial interpolation errors of high-resolution modeled atmospheric fields. *J. Atmos. Ocean. Tech*, 33, 303-311. doi: 10.1175/JTECH-D-15-0132.1

Putman, W., A.M. da Silva, L.E. Ott, and A.Darmenov, 2014. Model configuration for the 7-km GEOS-5 nature run, Ganymed release (non-hydrostatic 7-km global mesoscale simulation). *GMAO Office Note*. No. 5 (Version 1.0), Document (PDF, 6123 kB).

Tarantola, A., 1987. *Inverse problem theory. Methods for data fitting and model parameter estimation.* Amsterdam: Elsevier, 634 pp.

Timmermans, R.M.A., W.A. Lahoz, J.-L. Attié, V.-H. Peuch, R.L. Curier, D.P. Edwards, H.J. Eskes, P.J.H. Builtjes, 2015. Observing System Simulation Experiments for air quality. *Atmos. Environment*, **115**, 199-213. doi: 10.1016/j.atmosenv.2015.05.032