THE OSSE CHECKLIST

by Ross N. Hoffman and Robert Atlas

This document is a supplement to “Future Observing System Simulation Experiments,” by Ross N. Hoffman and Robert Atlas (Bull. Amer. Meteor. Soc., 97, 1601–1616) • ©2016 American Meteorological Society • Corresponding author: Ross N. Hoffman, Cooperative Institute for Marine and Atmospheric Studies, Rosenstiel School of Marine and Atmospheric Science, University of Miami, 4600 Rickenbacker Cswy., Key Biscayne, FL 33149 • E-mail: ross.n.hoffman@noaa.gov • DOI:10.1175/BAMS-D-15-00200.2

The observing system simulation experiment (OSSE) checklist presented here is meant to be a detailed guide to OSSE practitioners and OSSE consumers to evaluate the design of an OSSE system and OSSE experiments. It provides a step-by-step process designed to encourage the user to consider all aspects of good OSSE design principles. The current version of the OSSE checklist is directly applicable to weather-related observing, modeling, and data assimilation (DA) systems. But very similar issues and concerns apply to OSSEs that would be used across a diverse range of geophysical domains and the user of the OSSE checklist should be able to translate the discussion presented here to these domains fairly directly. Visit www.aoml.noaa.gov/qosap/osse-checklist/ for an interactive version of the OSSE checklist.

OVERALL.

✓ Is the OSSE system realistic?
  • Discussion: This is important for the results to be actionable.
  • Caution: If not, quantified results are not reliable, and even the sign of the impact may be questionable.

✓ Is the OSSE system appropriate to the impacts of interest?
  • Discussion: For example, if the object of the experiments is to investigate the impact of a new sensor on tropical cyclone (TC) intensity forecasting, this would not be possible if TCs do not develop realistically in the NR.
  • Caution: If not, results will not be reliable.

✓ Will the OSSE be completed quickly?
  • Discussion: It might take a very long time to design and execute an optimal OSSE. Some degree of nonoptimality may be acceptable.
  • Caution: Experimental results must be timely enough to support in decision-making.

✓ Are the caveats associated with the OSSE system well known?
  • Discussion: Given that no OSSE system is perfect, there will be some associated caveats. A particular OSSE system may be adequate for some investigations, but not others.
  • Caution: It is vital to delineate the limitations and deficiencies of the experiment and to not draw conclusions that go beyond that boundary.

YOUR EXPERIMENT.

This section is intended to record some minimal basic information about your OSSE. Answers here provide context for the questions that follow:

► What new instrument(s) and/or new technique(s) are you investigating?
What phenomena are of interest for the new instrument(s) and/or new technique(s)?

Describe your nature run (NR).

Describe your forecast and data assimilation (DA) systems.

Describe your experimental design (including your validation approach).

**Nature Run.**

**NR quality.**

✓ Is the NR sufficiently different from the forecast model?
  * Discussion: If the NR and forecast model are too similar there may be a twin problem, in which the DA system too easily converges on the truth.
  * Caution: If not, even with a small amount of data the analyses may be too good, leaving little room for improvement.

✓ Is the NR of sufficiently high resolution?
  * Discussion: The NR should well resolve phenomena of interest. Scales smaller than the observation footprint and/or the model resolution contribute to representativeness error.
  * Caution: If not, the phenomena of interest may not be realistic or even present in the NR. The observation simulation may be overly dependent on statistical parameterizations of the representativeness error.

✓ Is the NR sufficiently long?
  * Discussion: One year or more past spinup is desirable. A large ensemble of forecast cases provides greater statistical significance. Ideally, the OSSE will sample all relevant seasons.
  * Caution: If not, sample size may be small. Results may be representative of only a single season.

✓ Is the NR generated by an operationally vetted model?
  * Discussion: A great deal of effort is devoted to removing biases from operational models. For example, cloud amounts, locations, heights, and type should all be realistic, especially for simulating infrared (IR) satellite observations.
  * Caution: If not, the climatology of the NR may be different from reality.

✓ Is the NR a short- or medium-range forecast (or a collection of such forecasts)?
  * Discussion: This is a Quick OSSE, not a full OSSE. Quick OSSEs allow comparison to real and interesting forecast cases.
  * Caution: Generally, a single forecast is used as the NR in a Quick OSSE. This does not allow for actionable, statistically significant findings.

✓ Is the NR actually a reanalysis?
  * Discussion: As with the model-twin problem, a DA system can quickly converge to a reanalysis. Similarities are too great. New observing systems are handicapped, because in current data voids, the NR is basically a short-term forecast and therefore any simulated observations in these areas will not add new information.
  * Danger: If a reanalysis is used as the NR, then this is not really an OSSE; such an experiment cannot be used to evaluate the impact of new observing systems quantitatively.

**NR completeness.**

✓ Does the NR database save everything needed at sufficient temporal and spatial resolution?
  * Discussion: To accurately simulate observations, the highest possible temporal and spatial resolution is required.
  * Caution: If not, the simulated observations may not have realistic variability and representativeness errors.

✓ Are any required auxiliary data consistent with the NR?
  * Discussion: Examples might be aerosols, from land surfaces and volcanoes, sea ice extent, soil moisture, vegetative health, etc. These data sources should be consistent with the NR. For example, higher aerosol concentrations are expected downwind of source areas (such as the Sahara, the Aral Sea, etc.).
  * Caution: If not, simulated observations may have inconsistent error characteristics.

✓ Is there a capability to convert the NR data to initial conditions (ICs) for the forecast model?
  * Discussion: Forecast model ICs corresponding to the state of the NR are useful in some tests of the OSSE system.
  * Caution: If not, assimilation of perfect observations may be needed to generate NR ICs for the forecast model.
NR starting time.
✓ Are all the data types of interest for the experiments available at the starting time of the NR?
  • Discussion: One sensitive test of the simulated observations is to compare them to real observations at the start of the NR. Such comparisons of matching real and simulated observations are more powerful than statistical comparisons, which is the only possible approach later in the NR.
  • Caution: Simulated observations for systems that do not exist in reality at the start of the NR cannot be validated in this way. For example, simulated Infrared Atmospheric Sounding Interferometer (IASI) observations from a 2005 NR cannot be compared to real IASI observations in 2005, since the first IASI was launched in 2006.
✓ Is the beginning of the NR a good forecast of reality?
  • Discussion: For testing observation data at the start of the NR, the more similar the NR is to reality, the better. Since the exact start of the NR is arbitrary, a starting time should be selected that results in a very good forecast for the first few days.
  • Caution: If the start of the NR is a bust, only the very beginning can be used for one-to-one validation tests.

NR validation.
✓ Has the NR been properly validated?
  • Discussion: It is critical to thoroughly validate the NR.
  • Caution: An unvalidated NR may result in wasted effort and incorrect conclusions.
✓ Are the NR and real climate statistics similar? Are both annual and seasonal or ideally monthly statistics similar?
  • Discussion: To obtain similar data coverage and error characteristics, it is necessary to sample a climatology similar to reality. For example, realistic precipitation patterns are needed to obtain realistic patterns of data coverage and errors for microwave sensors. Comparisons should include at least latitude–height cross sections of the zonally and temporally averaged prognostic variables and maps of sea level pressure, integrated water vapor, 250-hPa winds, and 500-hPa heights.
  • Caution: If not, some parameters may need to be adjusted when simulating observations to ensure realistic patterns of data coverage and errors.
✓ Are the NR budgets balanced?
  • Discussion: If sea surface temperature (SST) is fixed, for example, the heat budget might not balance.
  • Caution: Unbalanced budgets may contribute to an unrealistic NR.
✓ Are the NR energetics (e.g., eddy kinetic energy (EKE)) on different space and time scales for different regions and seasons similar to reality?
  • Discussion: EKE is a proxy for variance of the atmospheric variables. If some scales have too little (much) variance in the NR, the analysis of these scales in the OSSE may be too good (poor).
  • Caution: If not, some phenomena of interest may not be well described in the NR or may not have realistic frequencies of occurrence.
✓ Is the NR representative of a typical period in reality?
  • Discussion: Consider an ensemble of a metric of interest from the past several years (or decades). Repeatedly replace one year selected randomly with a year from the NR. Is the NR an outlier? Does adding the NR noticeably alter the statistics of the ensemble?
  • Caution: The NR does not have to be identical in all statistical respects to reality, but if it cannot pass as a random draw from reality, the OSSE results may not be reliable.
✓ Are the phenomena of interest realistically depicted in the NR?
  • Discussion: For example, if storm (extratropical or tropical) forecasts are of interest, then storm track and intensity statistics should be similar to reality. Tracks in the NR should look reasonable compared to reality. For example, it would not be acceptable if all the cyclone tracks avoid North America. Track statistics should be compared by region and month to the real case. The statistics should include the number of cyclones, the number of cases of genesis and of cyclolysis, and the average central pressure, speed, and direction.
- Caution: If the statistics of a phenomenon of interest are skewed in the NR, then quantitative results from the OSSE will likely be skewed as well.

✓ Are the key variables needed to simulate the observations of interest realistically depicted in the NR?
  - Discussion: These key variables will typically include cloud amount, height, and type as well as sea level pressure, precipitation, and surface winds among others. For these variables plots of monthly and weekly averages, instantaneous fields should be examined.
  - Caution: If the distribution of variables needed to simulate the observations is different from reality, then the coverage and errors of the simulated observations will also be different from reality, unless additional measures are taken.

✓ Are the NR caveats documented?
  - Discussion: No NR is perfect. It is important to document any limitations that might affect the usefulness of the NR for different purposes and to not draw conclusions that go beyond those limitations.
  - Caution: Make certain that deficiencies in the NR do not eliminate the NR as applicable for the purposes of the OSSE.

**FORECAST MODEL.**

*Forecast model quality.*

✓ Is the forecast model a current operational model?
  - Discussion: To draw conclusions from the OSSE applicable to an operational model (or future planned operational model), then that model should be used in the OSSE system.
  - Caution: If not, results may not be applicable to forecast and DA systems of interest.

✓ Is the forecast model of sufficiently high resolution?
  - Discussion: In general, the forecast model resolutions should match the model resolution used in the system for which the OSSE results will be applied. In some situations, a lower-resolution version of an operational model may be adequate. A lower-resolution version is appropriate for testing and preliminary results and can be useful to narrow the list of interesting experiments that will then be conducted at full resolution.
  - Caution: If not, results may not be applicable at higher resolution.

✓ Are the phenomena of interest realistically depicted, maintained, and forecast by the model?
  - Discussion: There cannot be meaningful forecast impacts for phenomena of interest that are not realistically depicted, maintained, and forecast by the model.
  - Caution: If not, useful results cannot be obtained.

**Forecast model validation.**

✓ Are the climate statistics of the NR and forecast models similar?
  - Discussion: These differences may be identified as observational biases by the DA system.
  - Caution: Biases between the forecast model and NR that are different from biases between the forecast model and reality may skew the quantitative results of the OSSE compared to what would be obtained in reality.

✓ Are the NR and forecasts from the NR sufficiently different?
  - Discussion: Twin (i.e., predictability) tests should be conducted to determine this. Ideally, the differences between the NR model and the forecast models should be similar to the differences between a “state of the art” model and the real world. This can be tested by comparing forecasts from the NR model and the forecast model from the same (or very similar) initial conditions.
  - Caution: Forecast skill impacts must be interpreted carefully if the forecast skill in the OSSE is markedly different from forecast skill in reality.

**DATA ASSIMILATION (DA) SYSTEM.**

*DA system quality.*

✓ Is the DA system a current operational system?
  - Discussion: To draw conclusions from the OSSE applicable to an operational DA system (or future planned operational DA system), then that DA system should be used.
  - Caution: If not, results may not be applicable to DA systems of interest.

✓ Is the DA system of sufficiently high resolution?
  - Discussion: In general, the DA system resolutions should match the model resolution used in the system for which the OSSE results will be applied. In some situations, a lower-resolution version of an operational DA system may be adequate. A lower-resolution version is appropriate for testing and preliminary results and can be useful to narrow the list of interesting experiments that will then be conducted at full resolution.
  - Caution: If not, results may not be applicable at higher resolution.

✓ Are the phenomena of interest realistically depicted (analyzed) by the DA system?
  - Discussion: Problems would arise if the observations and short-term forecasts were of
good quality, but the DA overly smoothed or otherwise incorrectly analyzed a particular type of meteorological system.

- **Caution:** Conclusions about phenomena of interest will not be reliable if the analysis scheme is inappropriate for those phenomena.

**DA system completeness.**

✓ Are all DA system components activated in the OSSE?

- **Discussion:** These include variational bias correction (VarBC), all phases of the quality control (QC) system (observations should include some gross errors), including the variational QC, the vortex relocation, and bogus procedures [which would require simulated tropical cyclone vital observations (TCVitals) for tropical cyclones].

- **Caution:** Results from OSSEs that use an “incomplete” DA system may not be reliable because small changes in QC decisions can have large impacts on subsequent DA cycles.

**DA system validation.**

✓ Is the accuracy of the DA system in the OSSE similar to operational DA systems in reality?

- **Discussion:** A spinup experiment across the start of the NR can provide a sensitive comparison of the DA system in reality and in the OSSE.

- **Caution:** If not, careful interpretation of the results will be required.

✓ Does the DA system have similar sensitivity to observations in reality and in the OSSE?

- **Discussion:** Tools to assess this include comparing observing system experiment (OSE) to OSSE results for data denial of a current instrument and comparing the results of linearized assessment tools like forecast sensitivity to observations (FSO) in reality and in the OSSE control experiment.

- **Caution:** This is necessary for quantitative results to be reliable.

✓ Is the DA system properly calibrated?

- **Discussion:** A common calibration is to adjust the statistics of the added simulated errors as discussed below. Calibrating the simulated error statistics may not be sufficient. An a posteriori calibration using scaling determined from a parallel OSE/OSSE can be applied. Additionally, if the forecast model is too much like the NR model, the forecast model (or the NR model) could be modified by perturbing the constants used in the physical parameterizations (which of course are not known precisely) or by adding stochastic terms to the parameterization or directly to the model tendencies. Similar approaches have been taken to increase divergence among ensemble members in ensemble forecast systems.

- **Caution:** Quantitative results from an uncalibrated OSSE should be assumed to be unreliable.

✓ Are the percentages of observations removed by the QC procedures similar in reality and the OSSE?

- **Discussion:** QC summaries in the OSSE should be similar to those in reality.

- **Caution:** If the percentages differ, the observation simulation may be inaccurate, or the forecast model and the NR may be too similar.

**OBSERVATION SIMULATION.**

**Observation coverage.**

✓ Do the observations have realistic coverage in space, time, and, for radiances, spectral frequency for existing and planned sensors?

- **Discussion:** For example, if the NR is too cloudy, then cloud cover may need to be reduced for the purpose of simulating observations. Data coverage is known to have a significant impact on forecast skill. Plots of data coverage should be examined for each data type for at least one 6-h period in each season in the OSSE and in reality.

- **Caution:** Quantitative results may not be reliable if coverage is markedly different in the OSSE compared to reality.

**Observation operators.**

✓ Are the forward operators that convert the NR to observations realistic and accurate?

- **Discussion:** When the NR is converted to observation space, the results are referred to as “perfect observations.” However, the forward model creates some errors, which are part of the overall representativeness errors. The more closely the simulation procedures mirror the actual measurement process, the more realistic these representativeness errors will be.

- **Caution:** Simplistic forward operators may require more complex statistical models of observation error in order to get realistic representativeness errors and geophysical biases.

✓ Are the interpolation procedures accurate?

- **Discussion:** Temporal and spatial interpolation should be accurate. This should not be a problem if the NR is archived at full resolution.
Special attention should be paid to horizontal interpolation of surface properties that have substantial variation with elevation, such as surface pressure and surface temperature.

- Caution: Simplistic forward operators may require more complex statistical models of observation error.

✓ Are surface elevation differences between the NR, forecast model, and observation locations properly accounted for?
  - Discussion: In an operational system, differences between model terrain and actual observation elevation must be carefully accounted for. In an OSSE the situation is more complicated since the elevation of the NR terrain may be different than both the model terrain and the observation elevation.
  - Caution: Errors in handling surface elevation can greatly affect the value of the surface pressure and other surface-affected observations.

Observation errors.

✓ Are the radiative transfer (RT) forward models used to generate perfect observations different from those used in the DA system?
  - Discussion: Different RT models could be used for simulation and assimilation, or realistic uncertainty could be included by creating an alternative set of coefficients used in the fast RT Forward model for simulating observations in the OSSE while retaining the operational coefficients in the DA system. Aspects of the RT calculation used to generate the coefficients that might be perturbed include the line parameters, the treatment of physical processes, the spectral weighting function that corresponds to the sensor response, and the spectral resolution. Or the operational coefficients might simply be perturbed randomly.
  - Caution: Identical RT models eliminate an important error source that has the potential to produce biases in reality.

✓ Are atmospheric motion vector (AMV) and cloud track wind (CTW) observation locations and height assignment errors realistic?
  - Discussion: Two possible approaches are 1) to statistically estimate where trackable features should be found and the size of the height assignment errors and then use these to simulate AMVs and 2) to simulate imagery and then mirror the entire process of identifying features in the imagery, tracking these, and assigning heights—all following operational practice. Resolutions of 15 min and 2–5 km are used operationally and similar resolutions would be required in the NR to generate AMVs and CTWs from simulated imagery in an OSSE.
  - Caution: Current operational DA systems struggle with the complex error characteristics of AMVs and CTWs. If these types of errors are not realistic in the OSSE, then these observations may be “too good.”

✓ Are the equation of state and gravity model used to simulate GNSS/RO observations different from those used in the DA system?
  - Discussion: GNSS/RO observations as a function of altitude may be considered very precise observations traceable to absolute standards. However, the forward model used to determine model variables at the observation altitude depends on the determination of altitude hydrostatically through the equation of state and the gravity model. As with the discussion of the RT models above, there should be some differences between the observation simulation and the forward model used in the DA system.
  - Caution: Identical equations of state and gravity models eliminate an important error source that has the potential to produce biases in reality.

✓ Are there appropriate differences between the surface characteristics used to simulate data and used to assimilate those data?
  - Discussion: This includes temporal changes of soil moisture, sea ice, and vegetation. As with the discussion of the RT models above, there should be some differences between the observation simulation and the forward model used in the DA system.
  - Caution: Identical surface specifications eliminate an important error that has the potential to produce biases in reality.

✓ Are added observational errors realistic?
  - Discussion: Explicitly added observational errors should have the right characteristics including biases, error–error correlations (vertical for raobs, horizontal for CTWs, channel for radiances), and error–geophysical parameter correlations. These geophysical parameters include the synoptic situation and land surface properties. The amplitude of the error variances and error length scales may also depend on such geophysical parameters. For example, errors may be larger in cloudy, precipitating, and/or stormy conditions.
Caution: The correct specification of the explicitly added observational errors is critical to a well-calibrated OSSE system.

EVALUATION STRATEGY.
Quantification and interpretation of impacts and results.
✓ Are the impact metrics appropriate for the phenomenon of interest?
  • Discussion: Anomaly correlation coefficients of 500-hPa geopotential and vector wind root-mean-square errors at 200 and 850 hPa are traditional metrics used to evaluate forecast skill. Evaluation should usually include other metrics, perhaps subjective assessment of forecast skill and objective methods that separate position error from other errors.
  • Caution: Results are difficult to interpret if the metrics are not appropriate for the phenomena of interest.
✓ For weather forecasting, are synoptic impacts examined as well as statistical skill score impacts?
  • Discussion: In many cases the sample of interesting cases of forecasts of phenomena of interest may be limited. If so, statistical evaluations may be inconclusive. A physical understanding through a synoptic evaluation of the impact in the OSSE can provide more insight and reliability.
  • Caution: Without a synoptic evaluation, important impacts may be hidden, and understanding of the value of the new observations or method may be incorrect.
✓ Are probabilistic ensemble forecast metrics included in the evaluation?
  • Discussion: There is an increasing emphasis on providing probabilistic forecast information through the use of ensembles. Large samples are needed to validate probabilistic forecasts.
  • Caution: If such metrics are not included, the OSSE cannot reveal any impact on probabilistic forecasts.
✓ In addition to the full OSSE analysis, is FSO used to determine if the new observation might make some existing observations redundant?
  • Discussion: If the DA system includes components that provide or allow linear sensitivity analyses, these should be used. FSO and related techniques can provide insights into which observing systems are redundant or partially redundant to a new observing system.
  • Caution: Without FSO it may not be apparent that the impact of the new observing system is being masked by one of the existing observing systems. In some situations, this might be valuable in interpreting the OSSE results.
✓ Does the evaluation strategy include providing levels of statistical significance?
  • Discussion: Decision-makers and others reviewing the reported OSSE results will want to know if the impacts quantified are statistically significant. A prerequisite is that the OSSE period must be sufficiently long to permit statistically significant results.
  • Caution: Without a level of significance, quantitative OSSE results may be difficult to use in subsequent decision-support analysis.