Observing System Simulation Experiments

30 July 2015

Nikki Privé

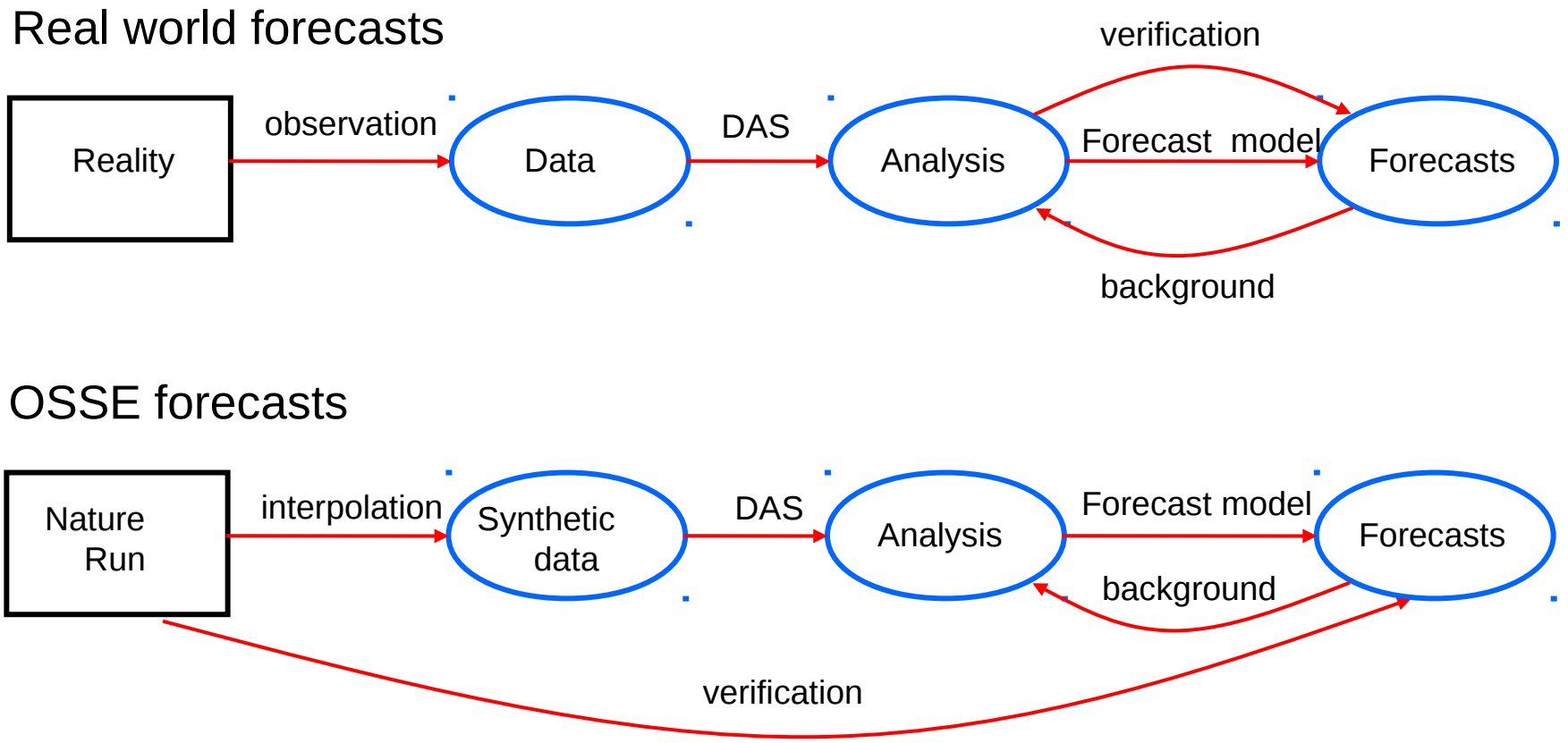
Ron Errico

What is an OSSE?

An OSSE is a modeling experiment used to evaluate the impact of new observing systems on operational forecasts when actual observational data is not available.

- A long free model run is used as the “truth” - the Nature Run
- The Nature Run fields are used to back out “synthetic observations” from all current and new observing systems.
- Suitable errors are added to the synthetic observations
- The synthetic observations are assimilated into a different operational model
- Forecasts are made with the second model and compared with the Nature Run to quantify improvements due to the new observing system

OSSEs vs. the Real World



Why do an OSSE?

1. You want to find out if a new observing system will add value to NWP analyses and forecasts
2. You want to make design decisions for a new observing system
3. You want to investigate the behavior of data assimilation systems in an environment where the truth is known

When not to run an OSSE

- When you can't model the phenomena you are interested in
- When you can't simulate your new observations
- When you can't assimilate your new observations

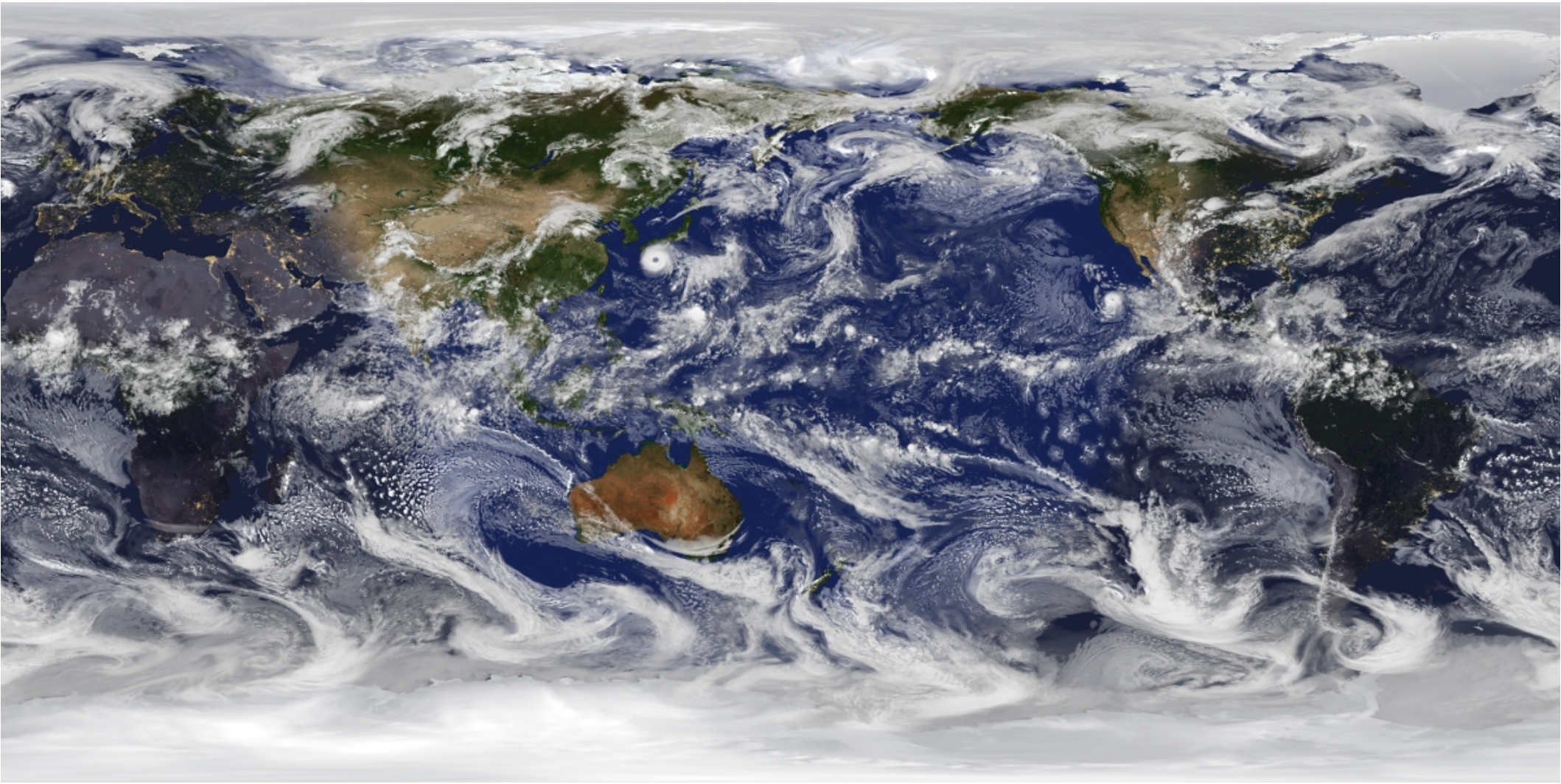
Nature Runs

- Nature Runs act as the 'truth' in the OSSE, replacing the real atmosphere.
- Usually, a long free (non-cycling) forecast from the best available model is used as the NR
  - Model forecast has continuity of fields in time
  - Sometimes an analysis or reanalysis sequence is used, but the sequence of states of truth can never be replicated by a model
  - Always a push for bigger, higher resolution NR

Nature Run Requirements

- Must be able to realistically model phenomena of interest
  - Dynamics and physics should be realistic
  - Must produce fields needed for “observations”
  - Should be verified against real world
- Ideally is ‘better’ than the operational model to be used for experiments
- Preferably a different model base is used for the NR and the experimental forecast model to reduce incestuousness

G5 Nature Run

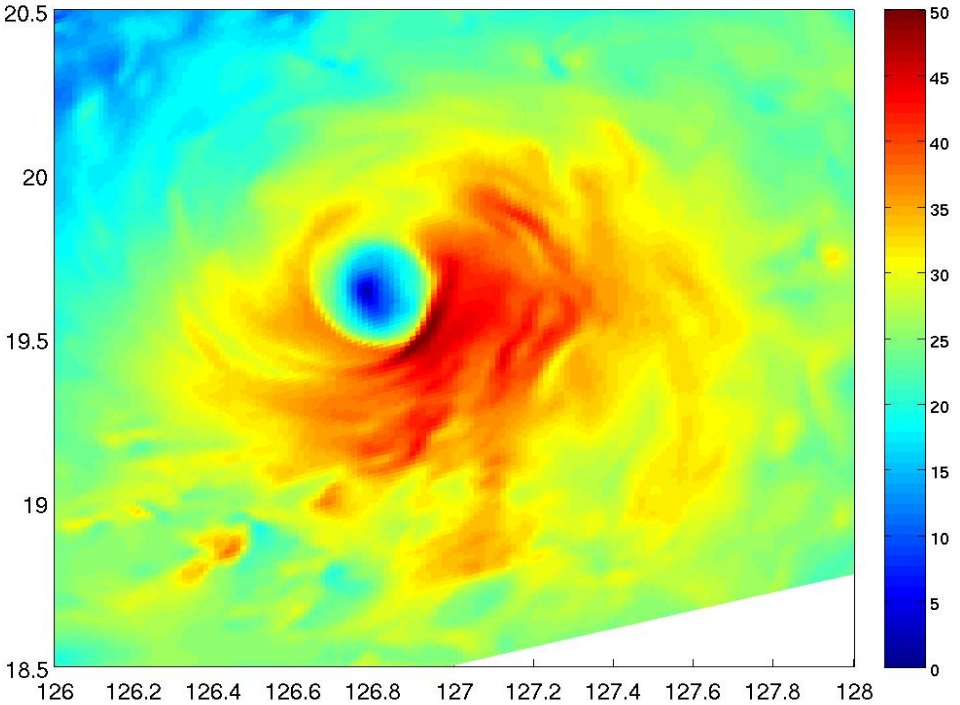


2 year, 7 km/72L, 30 minute resolution
15 aerosols, ozone, CO, CO_2

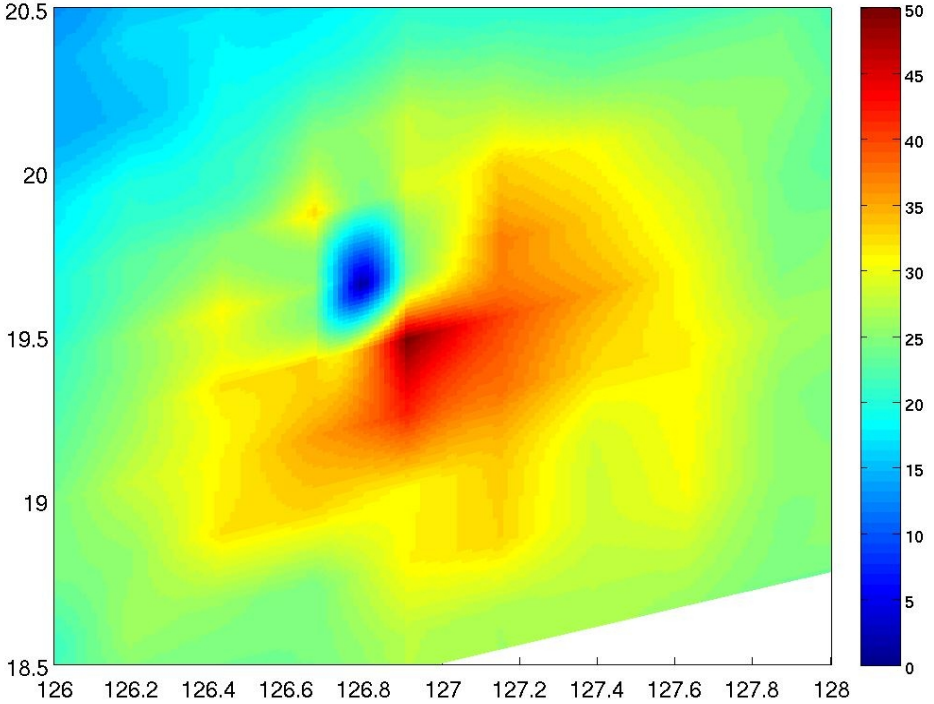
Common Problems with Nature Runs

- Nonexistence
- Identical or fraternal twins
- Outdated by the time you get to use them
- Gigantic output files and huge computational resource requirements
  - Output saved at full spatial resolution but 30 min + intervals

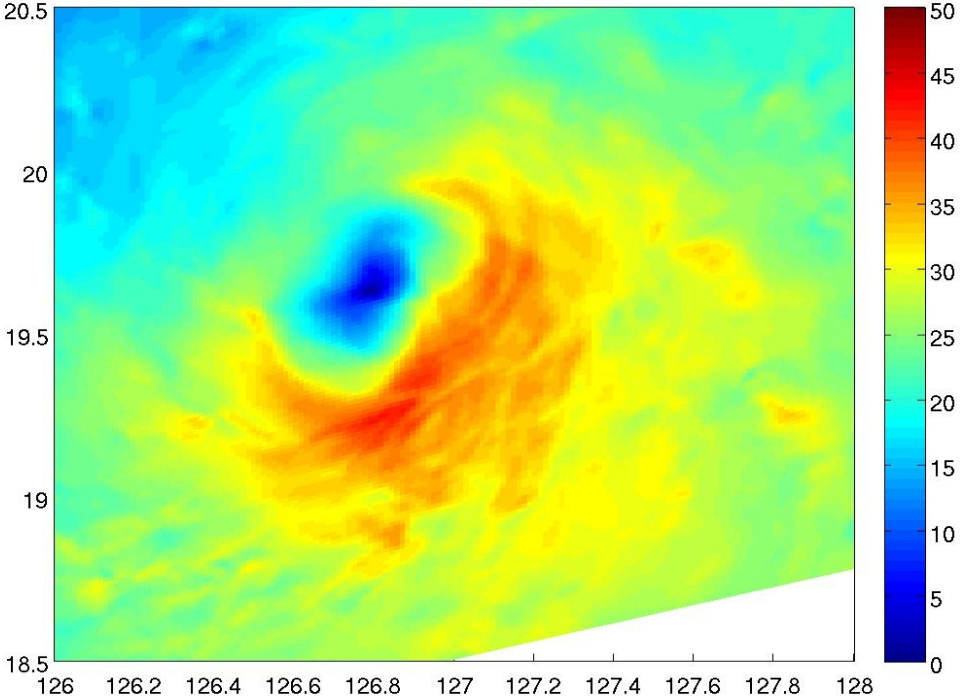
Full resolution (1.5 km, 10 min)



Spatial interpolation to 27 km



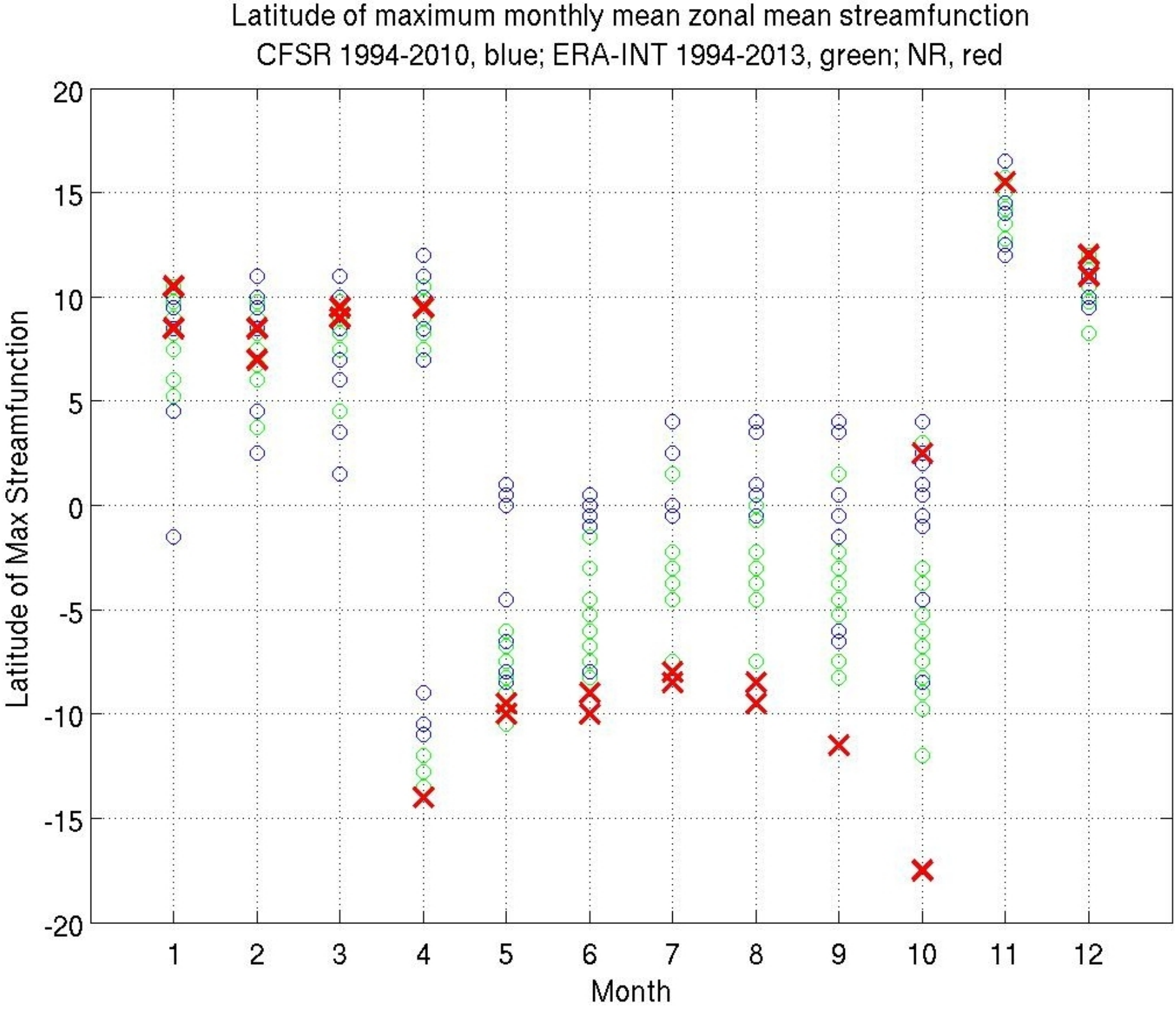
Temporal interpolation (3 hrs)



Comparison of temporal and spatial Interpolation errors compared to 1.5 km run for Typhoon Guchol (2012).

Nature Run Validation

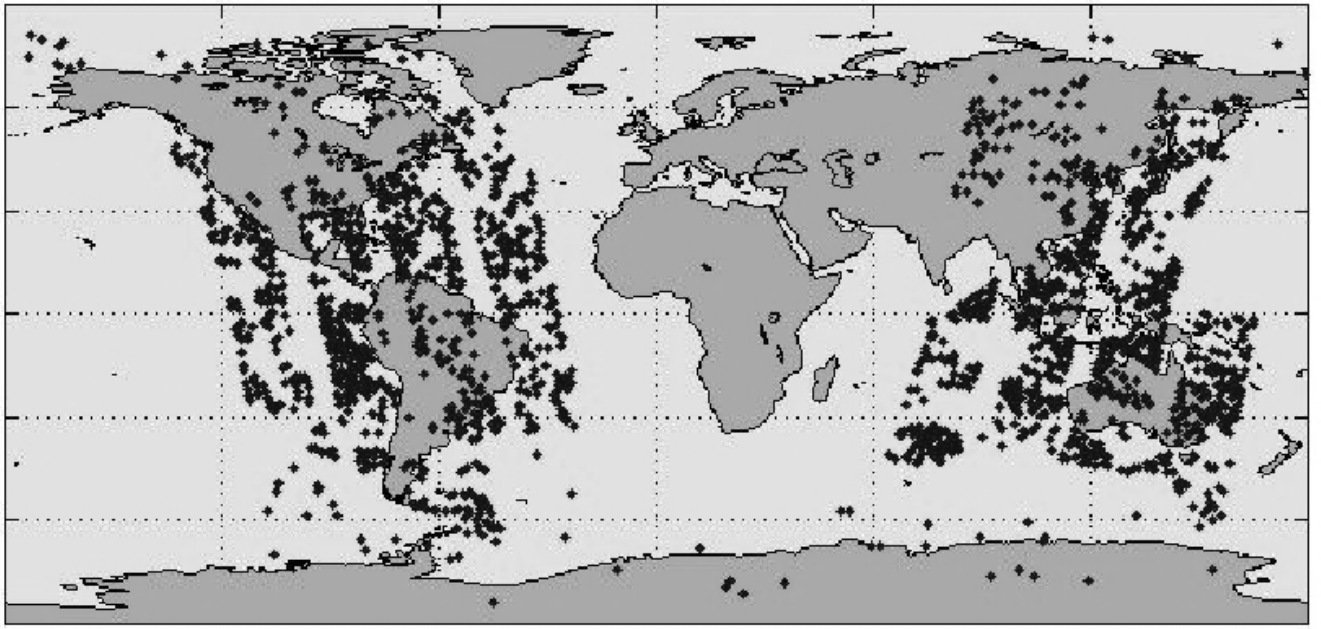
- Evaluate if NR is sufficiently realistic to yield meaningful results
- In addition to the phenomena of interest, the NR needs to realistically replicate fields needed to generate synthetic observations
- Can't validate everything; corollary – don't expect a NR to come pre-validated for your needs


Latitude of maximum monthly mean zonal mean streamfunction
CFSR 1994-2010, blue; ERA-INT 1994-2013, green; NR, red
Latitude of Max Streamfunction
Month
20
15
10
5
0
-5
-10
-15
-20
1
2
3
4
5
6
7
8
9
10
11
12

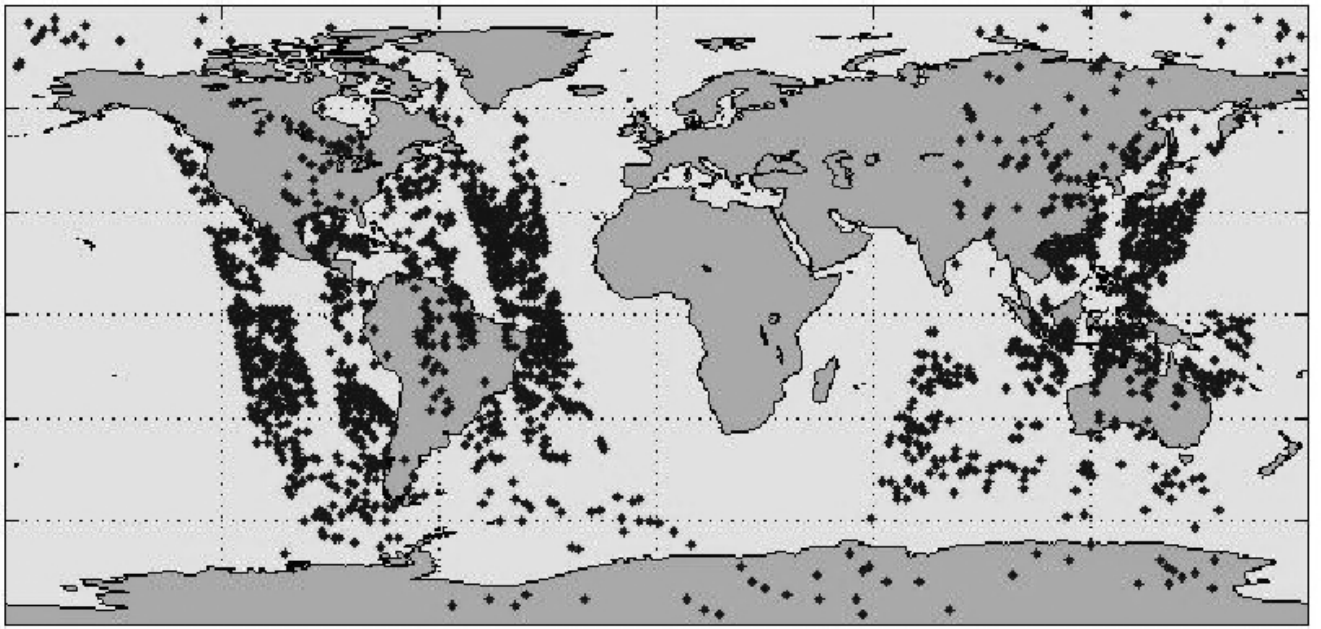
Synthetic Observations

Example of AIRS observations channel 295 at 18 UTC 12 July

Simulated



Real

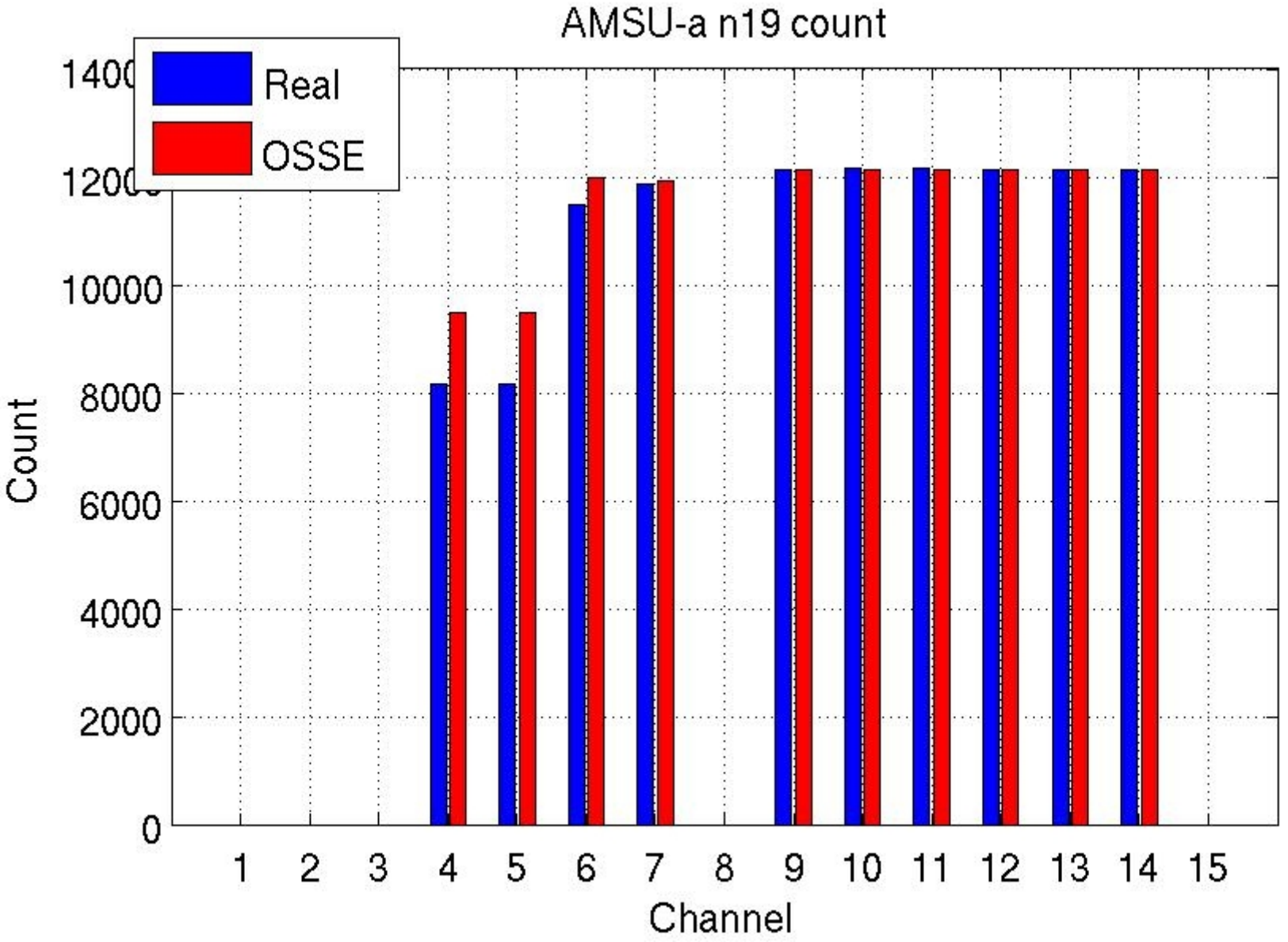


Observation Errors

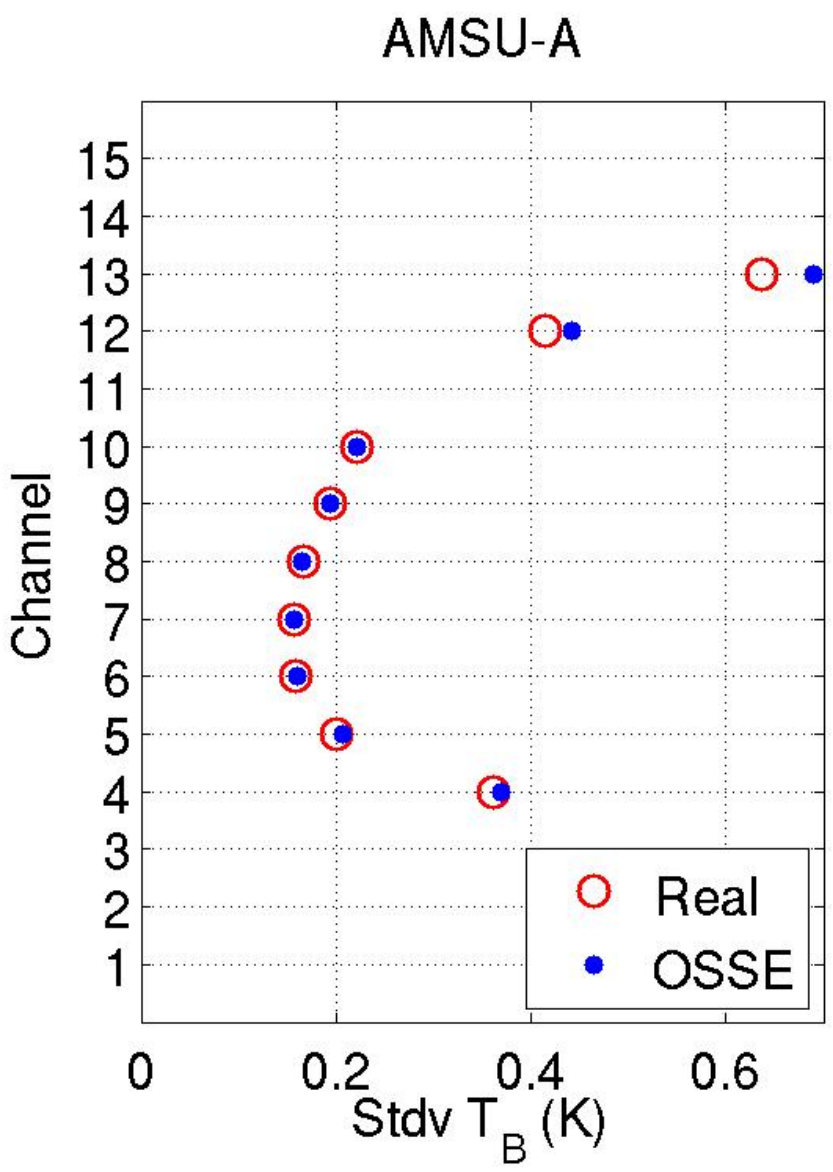
- Synthetic observations contain some intrinsic interpolation/operator errors, but less than real observations (usually)
- Synthetic errors are created and added to the synthetic observations to compensate
- Error is complex and poorly understood
  - Error magnitude
  - Biases
  - Correlated errors

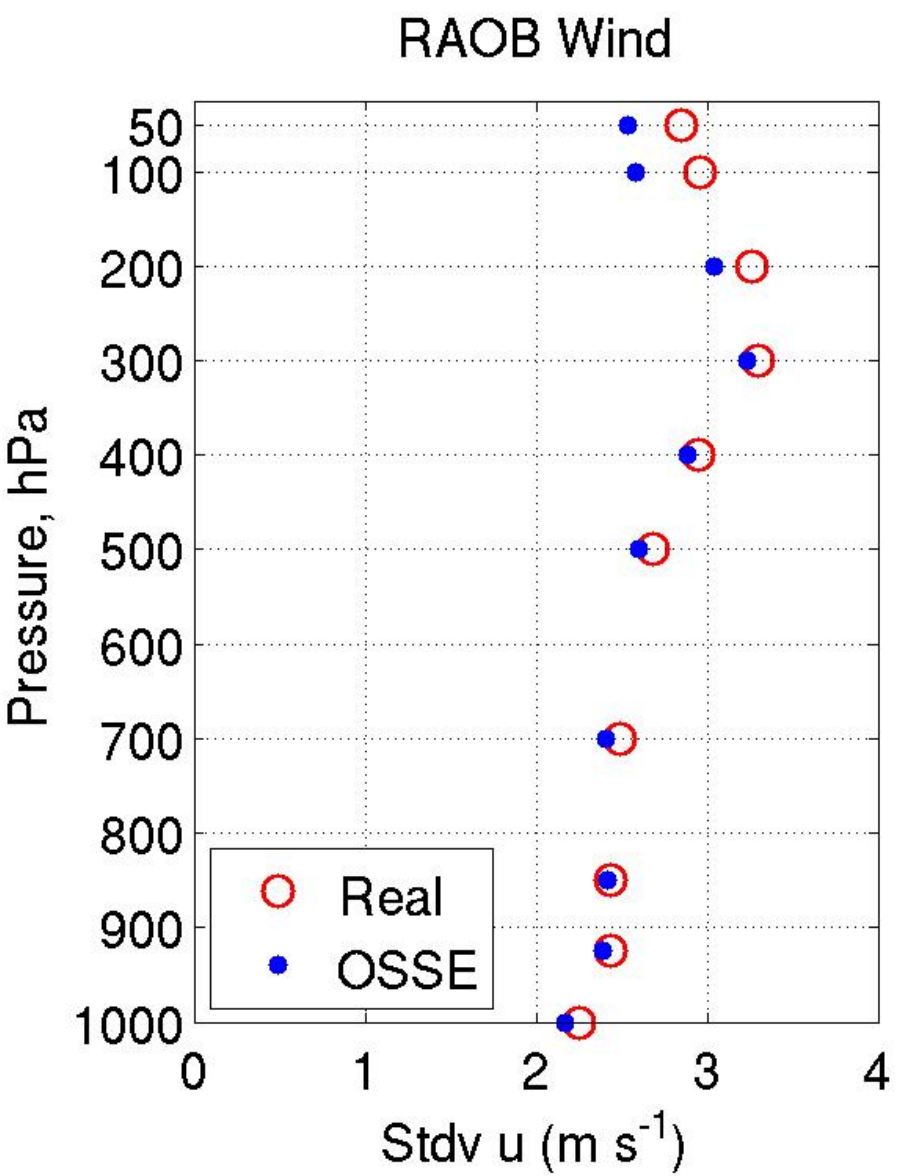
Calibration

- Adjust synthetic observations and their errors to increase realism of the OSSE in a statistical sense
  - Compare OSSE statistics to statistics using real data in the same DAS/forecast system
- Need to decide what statistical metrics to use for the calibration, depending on your needs
- Calibrating new observation types?
  - Find an analogous data type if possible

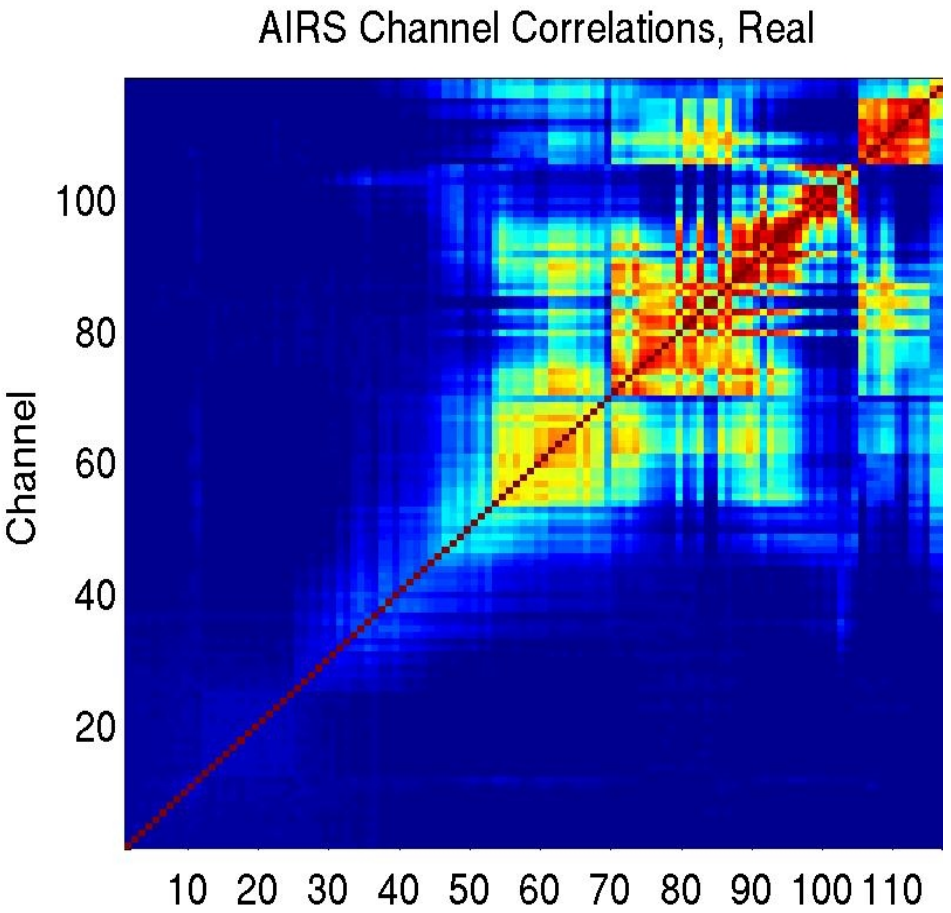


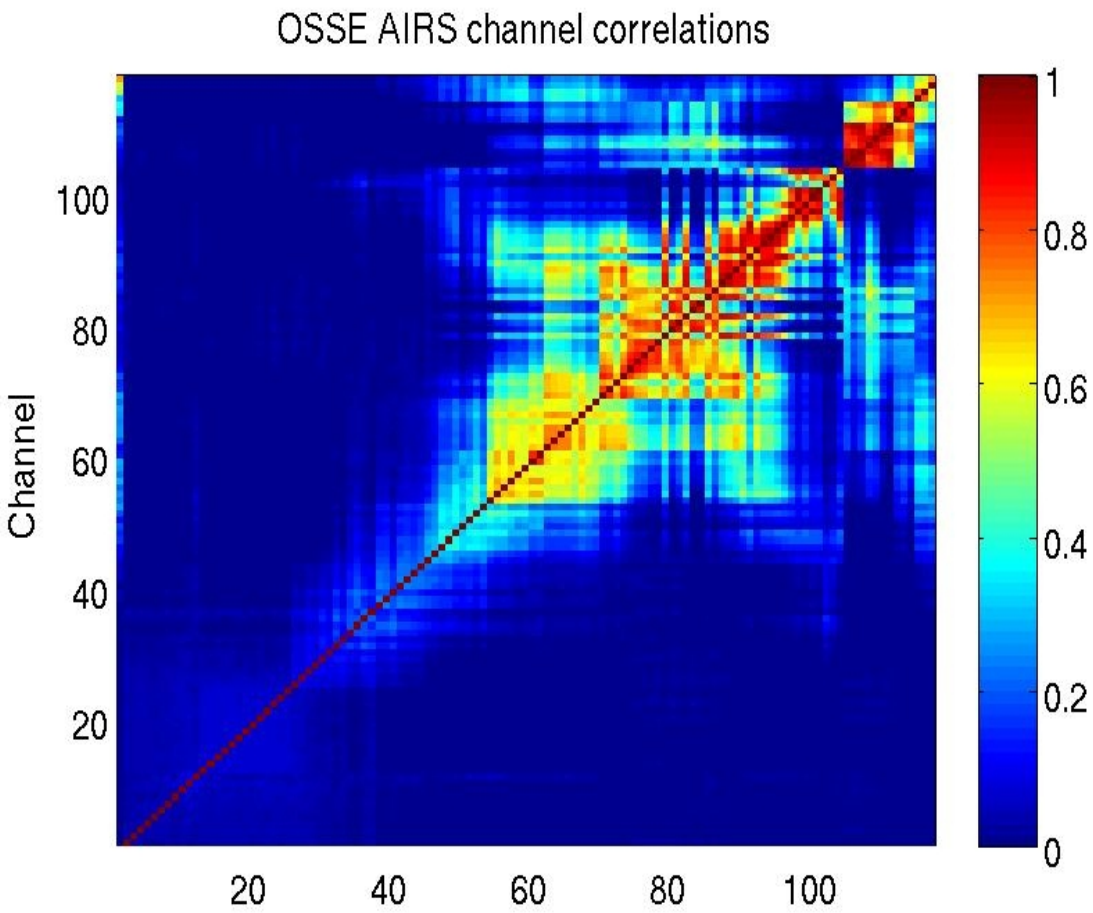
Observation count is easy to calibrate



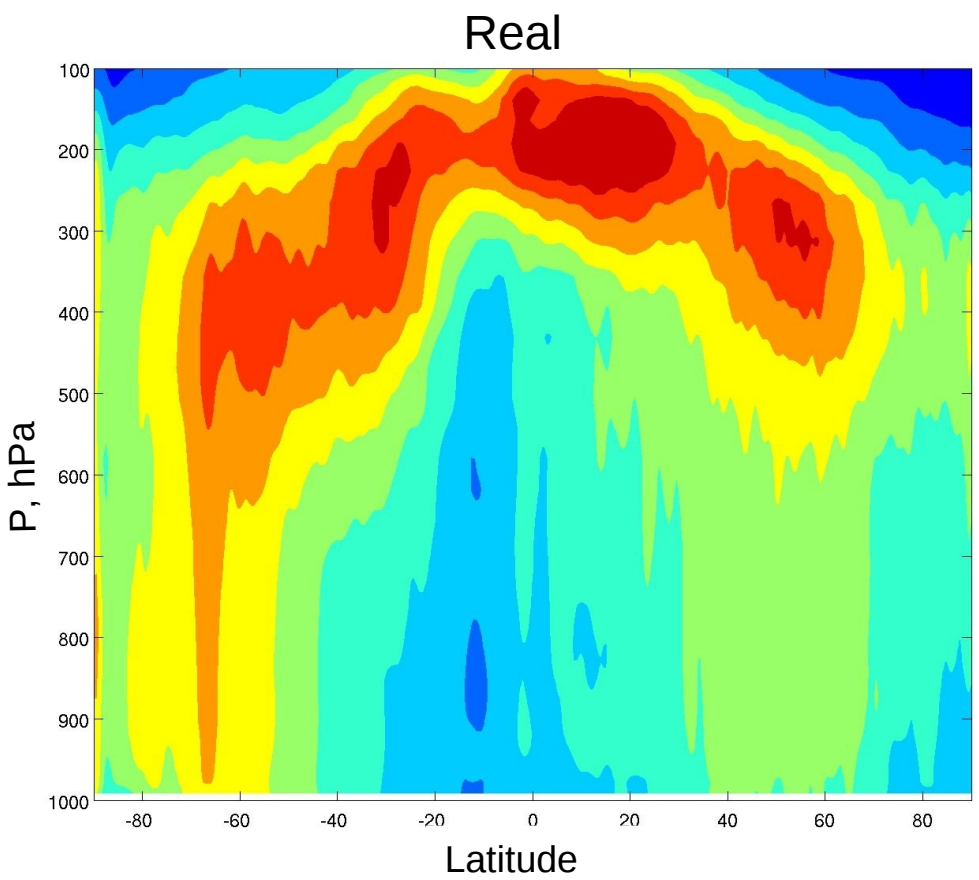


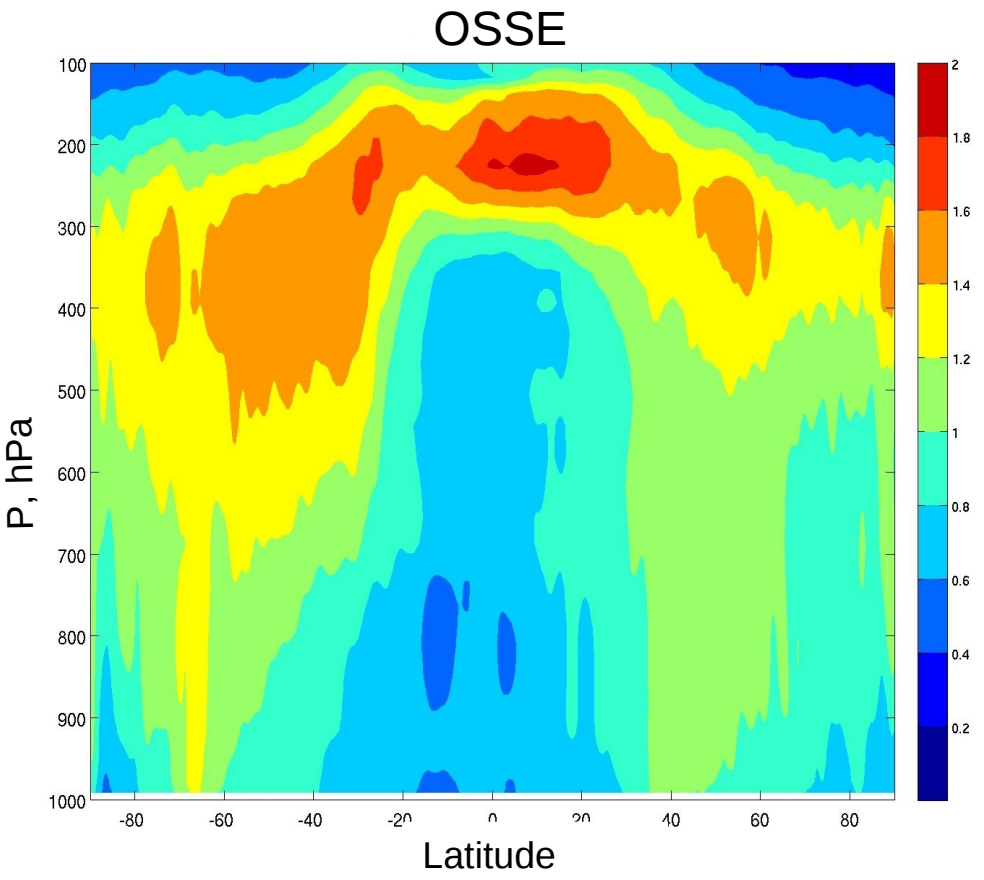
O-F is fairly easy to calibrate because you can manipulate O directly.



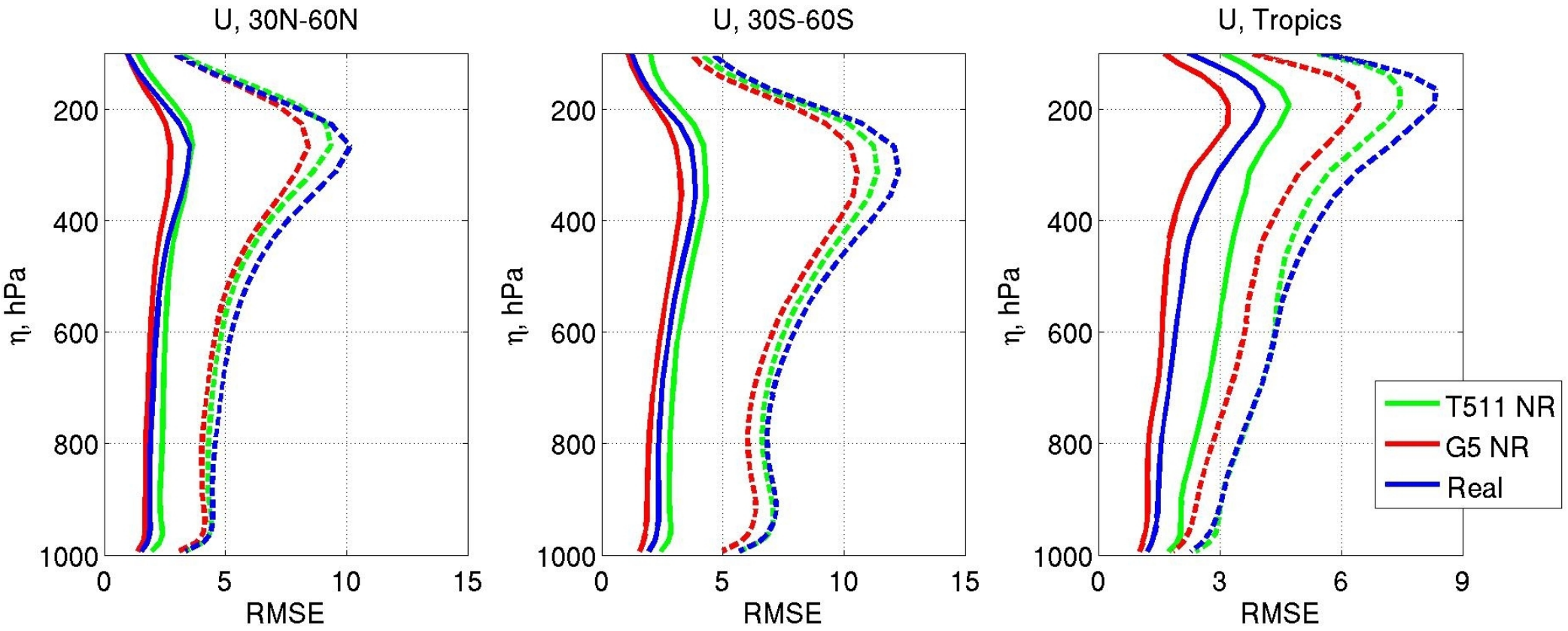


Some observation correlations are relatively easy to calibrate





A-B (analysis increment) is a little harder to calibrate, as A and B are not directly controlled

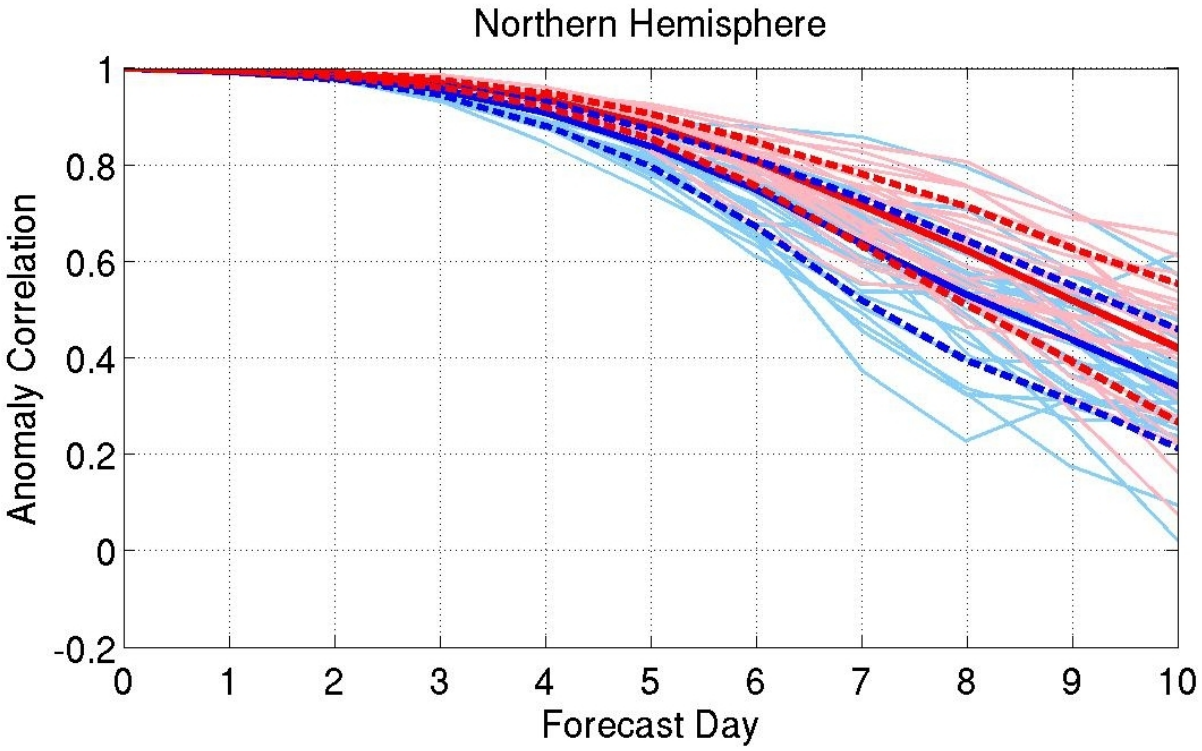


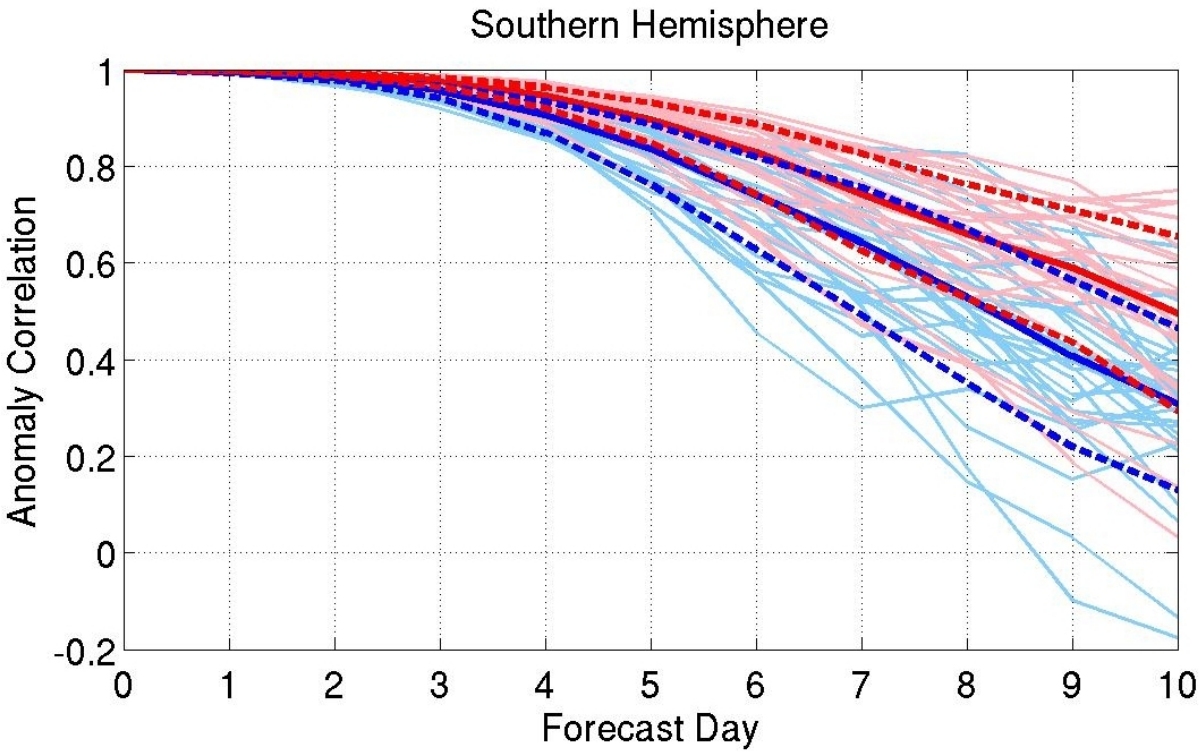
Forecast errors are harder to calibrate, especially for longer forecasts. Matching of this statistic by manipulation of observations is difficult to impossible beyond ~24 hour forecasts.

Model error determines forecast skill in the longer term forecast, so calibration is not possible (unless you want to mess with your model).

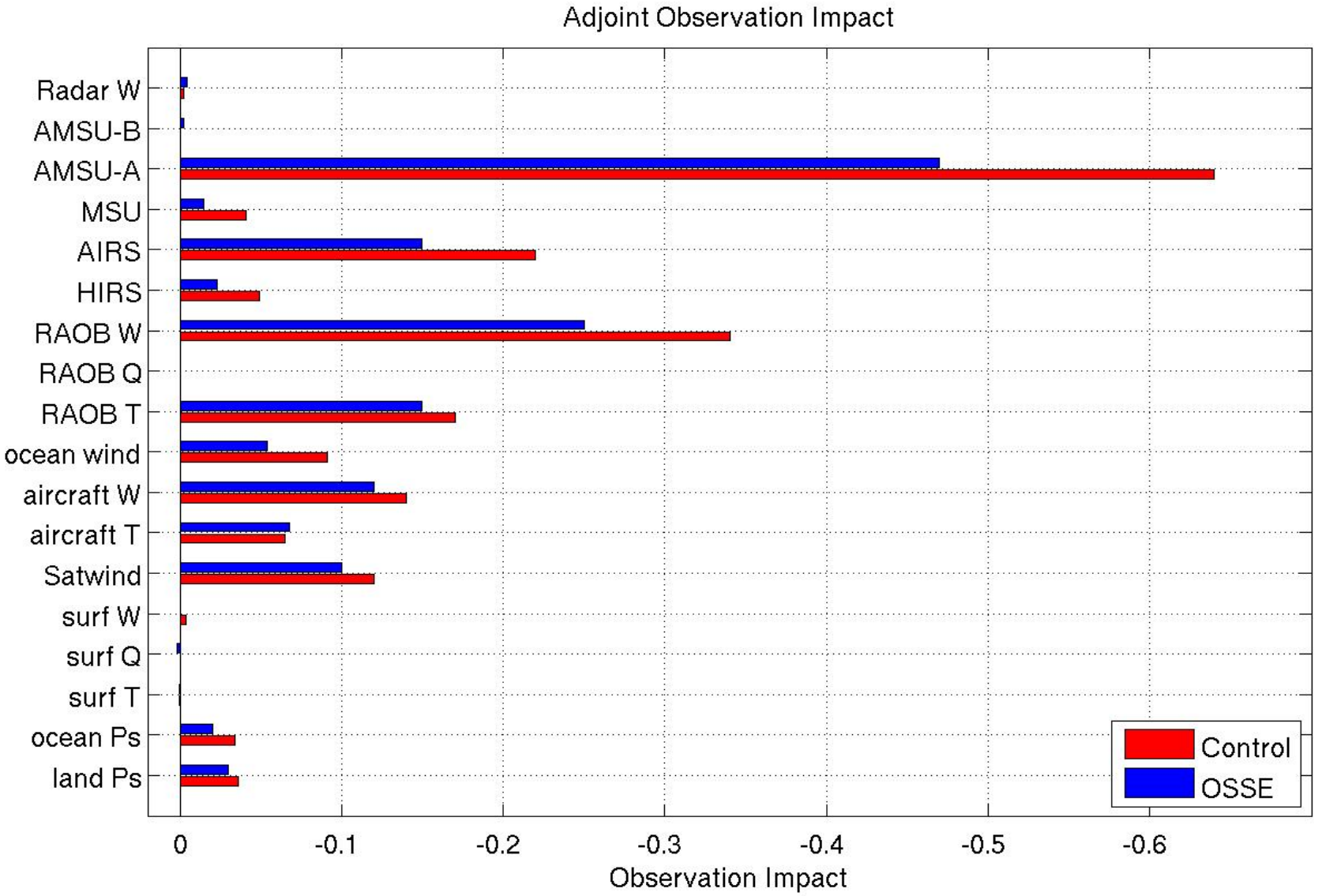
Red: OSSE
Blue: Real

500 hPa anomaly correlations of geopotential height





Why believe OSSE results?



New observations can be put into context relative to existing observation impacts

Criticisms of OSSEs

- Results only apply within the OSSE system – no concrete connection to the real world
- Even the best OSSEs are far from perfect: incestuousness, difficulty in generating observations and errors, deficiencies of the Nature Run
- By the time the new instrument is deployed, both the global observing network and the forecast models/DAS will be different
- Examples of sloppy or unsuccessful OSSEs

Common Pitfalls

- Very reduced baseline of assimilated observational data (ex. no radiance data)
- Other artificial degradation of analysis state
- No validation or calibration of OSSE framework
- Obtaining robust results from case studies is very challenging
  - Use ensemble forecasts if you can!

Choosing Metrics

- Long cycling periods necessary to get statistically significant results for most new observations
- Anomaly correlation is a difficult metric to show appreciable impacts
- What fields do you expect the instrument to improve?
- Largest impacts found at analysis time or short-term forecasts

Idealized Studies

- Identical twin experiments
- Idealized observations
- Manipulation of observation errors
- Experiments with **B**, **R**

- Make use of available “Truth”

Regional OSSEs

- Regional OSSEs are harder than global OSSEs
  - Two Nature Runs (local embedded in global)
  - Two forecast models each using synthetic obs
  - Shortcuts are very often taken – need to carefully examine the methods to ascertain if the results are trustworthy

Takeaway

- OSSEs can provide useful information about new observational types and the workings of data assimilation systems
- Careful consideration of research goals should guide each step of the OSSE process
- OSSEs are hard, good OSSEs are harder

References

Review article: 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. Environ., **115**, 199-213.

Observation errors: 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

Regional OSSEs: Nolan, David S., Robert Atlas, Kieran T. Bhatia, and Lisa R. Bucci, 2013: Development and validation of a hurricane nature run using the Joint OSSE Nature Run and the WRF model. J. Adv. Earth. Model. Syst., **5** , 1-24

Early history of OSSEs: Arnold, C.P. And C.H. Dey, 1986: Observing System Simulation Experiments: past, present, and future. Bull. Amer. Meteor. Soc., **67**, 687-695.

Observing System Simulation Experiments

30 July 2015

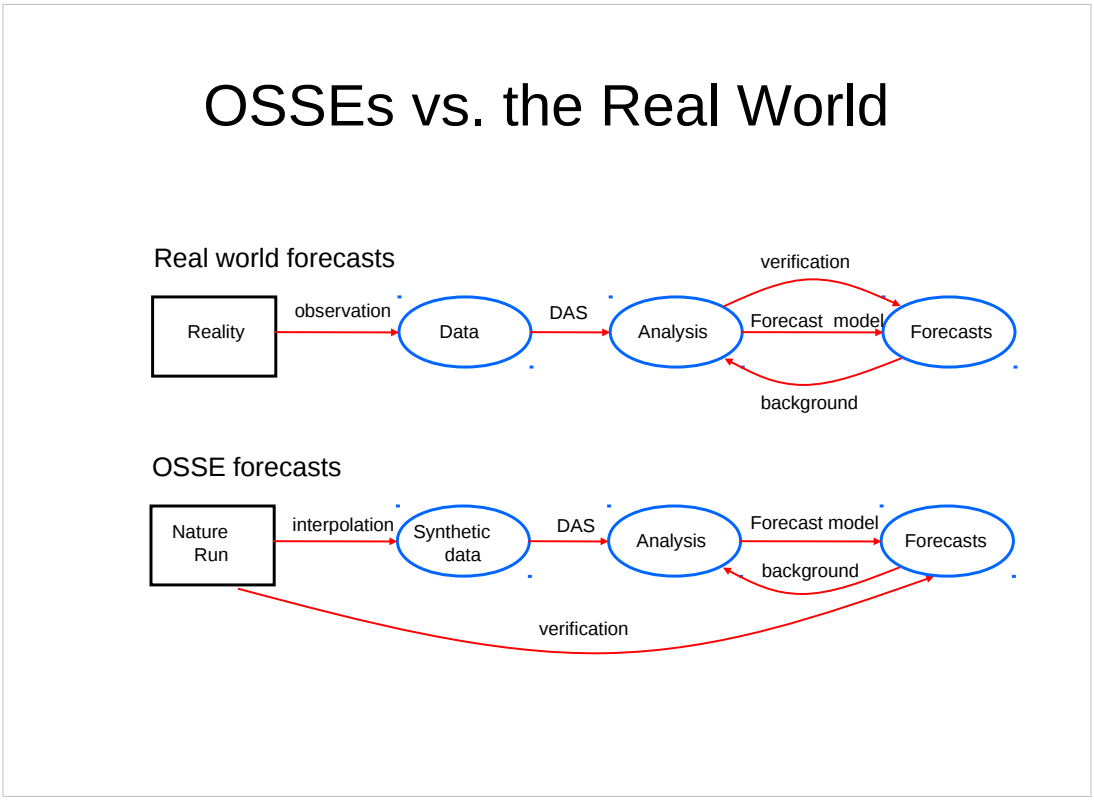
Nikki Privé
Ron Errico

What is an OSSE?

An OSSE is a modeling experiment used to evaluate the impact of new observing systems on operational forecasts when actual observational data is not available.

- A long free model run is used as the “truth” - the Nature Run
- The Nature Run fields are used to back out “synthetic observations” from all current and new observing systems.
- Suitable errors are added to the synthetic observations
- The synthetic observations are assimilated into a different operational model
- Forecasts are made with the second model and compared with the Nature Run to quantify improvements due to the new observing system

Basic components of the `traditional' OSSE used to evaluate new observation impact in a numerical weather prediction framework. Other uses for OSSEs include investigation of the behavior of data assimilation systems and `climate OSSEs' (which have some similarity to NWP OSSEs but also some major structural differences and will not be discussed further).



Flowchart comparing OSSEs to operational data assimilation. The two main differences are the replacement of the actual atmosphere with the Nature Run (NR), and the use of the NR to verify forecasts (and analyses) in the OSSE instead of self-analysis verification.

Why do an OSSE?

1. You want to find out if a new observing system will add value to NWP analyses and forecasts
2. You want to make design decisions for a new observing system
3. You want to investigate the behavior of data assimilation systems in an environment where the truth is known

The “classic OSSE” is the first item on this list – testing new observing system designs in a controlled framework when actual observations are not available (instrument doesn't exist, too expensive, etc).

A potentially valuable study is to compare design possibilities for a new instrument – for example, decisions about scan characteristics, orbits, channels, for a new satellite. This can be used to determine if a less expensive alternative design will still yield useful results in comparison to the full design.

The third point is valuable in the context of design and understanding of data assimilation systems. In the OSSE, the availability of the truth for verification allows the direct calculation of analysis errors, as well as adjoint and other observation impact metrics, to determine exactly how the ingestion of different observations affects the skill of the analysis. These calculations are not possible in the real world environment.

When not to run an OSSE

- When you can't model the phenomena you are interested in
- When you can't simulate your new observations
- When you can't assimilate your new observations

1. If you are interested in something that your available models (BOTH NR and forecast model) produce either poorly or not at all (due to resolution, improper physics, etc), you may be wasting your time with an OSSE. In some cases, you might be able to find an alternative field that the model does reproduce well, and which has a link to the actual phenomenon of interest.

2. If your NR does not produce the output needed to simulate your observations, or if you just don't know how to simulate the observations well enough, do not do an OSSE.

3. If your desired forecast model does not have suitable DAS to go along with it, or if your observations are not suitable for assimilation into NWP models, do not do an OSSE.

Nature Runs

- Nature Runs act as the 'truth' in the OSSE, replacing the real atmosphere.
- Usually, a long free (non-cycling) forecast from the best available model is used as the NR
  - Model forecast has continuity of fields in time
  - Sometimes an analysis or reanalysis sequence is used, but the sequence of states of truth can never be replicated by a model
  - Always a push for bigger, higher resolution NR

Both re/analysis based and model based NR are found in the literature. The model-based NR are often considered preferable to analyses/reanalyses.

Nature Run Requirements

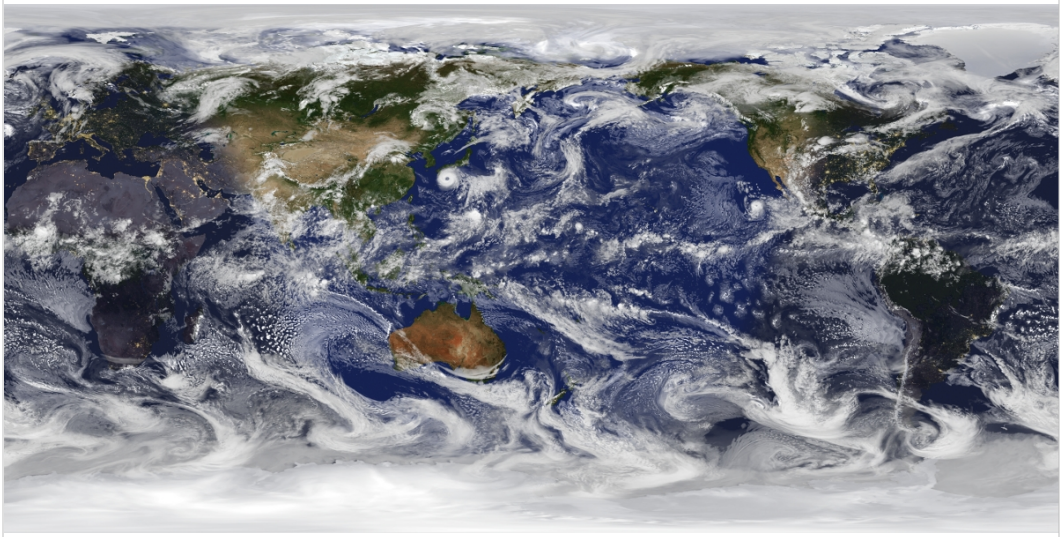
- Must be able to realistically model phenomena of interest
  - Dynamics and physics should be realistic
  - Must produce fields needed for “observations”
  - Should be verified against real world
- Ideally is ‘better’ than the operational model to be used for experiments
- Preferably a different model base is used for the NR and the experimental forecast model to reduce incestuousness

NR should ideally be able to replicate the atmosphere realistically for not just the phenomena that you are interested in forecasting, but also everything needed to simulate all of the observations used by the DAS. So, AMVs need a realistic distribution of clouds, etc.

The NR preferably should have higher spatial resolution than the forecast model, and also be more realistic in terms of dynamics/physics. In practice, the comparative resolution of the NR and the forecast model is often rather similar due to both available resources and limitations of model development. When NR are made with bleeding-edge high resolution models, the model behavior can be rather buggy, with undesirable behaviors.

In general, the model used to made the NR and the model used for the experimental forecasts are more similar to each other than either model is to the real world. Thus, OSSEs tend to suffer from insufficient model errors. This is more true for “identical twin” cases in which the same model is used for the NR and the experimental forecasts. Lack of model error has complicated impacts, as the model forecast skill tends to be too high, resulting in too-skillful backgrounds, so observations have less impact, but the information imparted by the observations may have a better chance of 'surviving' into the extended forecast instead of being destroyed by model error.

G5 Nature Run



2 year, 7 km/72L, 30 minute resolution
15 aerosols, ozone, CO, CO2

The new GMAO Nature Run is now available. 7 km horizontal resolution (C5760 cube sphere native), 72 levels, includes aerosols but not most chemistry. 24 months using SSTs from May 2005-May 2007.

The NR data may be accessed from http://gmao.gsfc.nasa.gov/projects/G5NR/

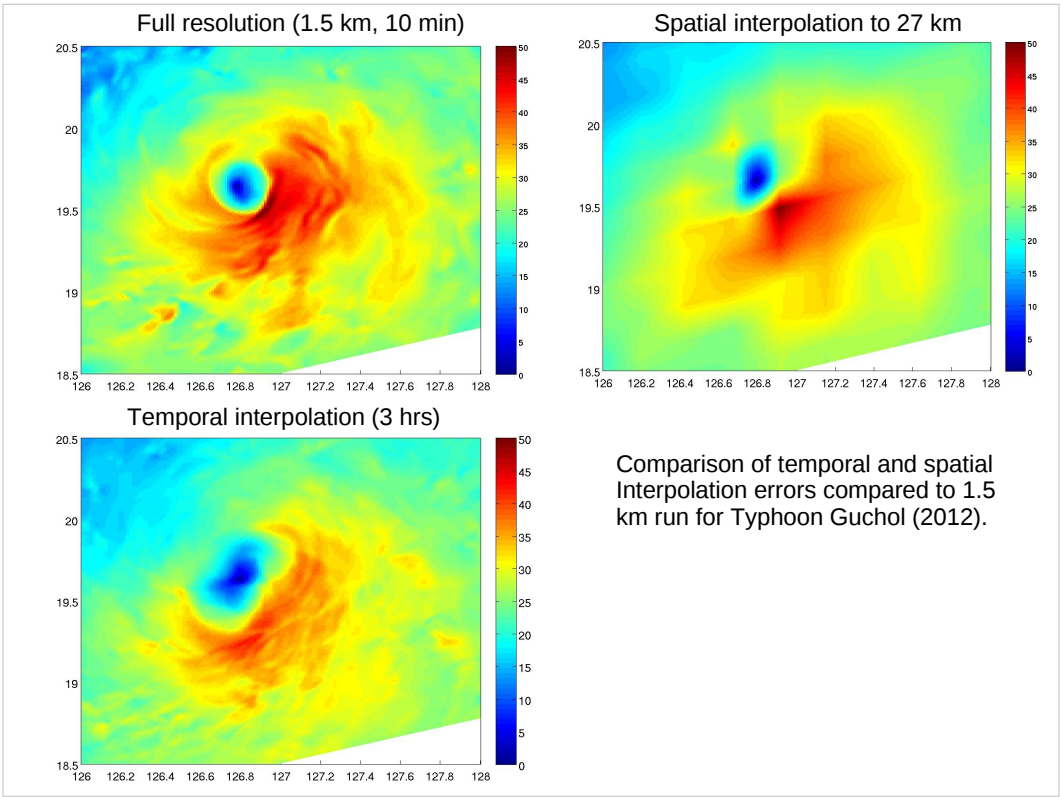
Validation of the G5NR is documented in this technical memo:
http://gmao.gsfc.nasa.gov/pubs/docs/TM2014-104606v36.pdf

Common Problems with Nature Runs

- Nonexistence
- Identical or fraternal twins
- Outdated by the time you get to use them
- Gigantic output files and huge computational resource requirements
  - Output saved at full spatial resolution but 30 min + intervals

Many different approaches are taken in the literature – some good, some not so much. Identical and fraternal twin OSSEs are fairly common, and not always a terrible thing depending on how the OSSE is performed. In some “OSSE”s, a true Nature Run is not used in that observations are not simulated using the NR fields – these types of studies can be fraught.

Generation of NR s can be very computationally expensive. The output files can be so large that they are difficult to handle with standard software. The length of time that it takes to create and validate a NR can mean that by the time the NR is ready to use, operational forecast models have already caught up to the resolution and physics used in the NR (or surpassed).



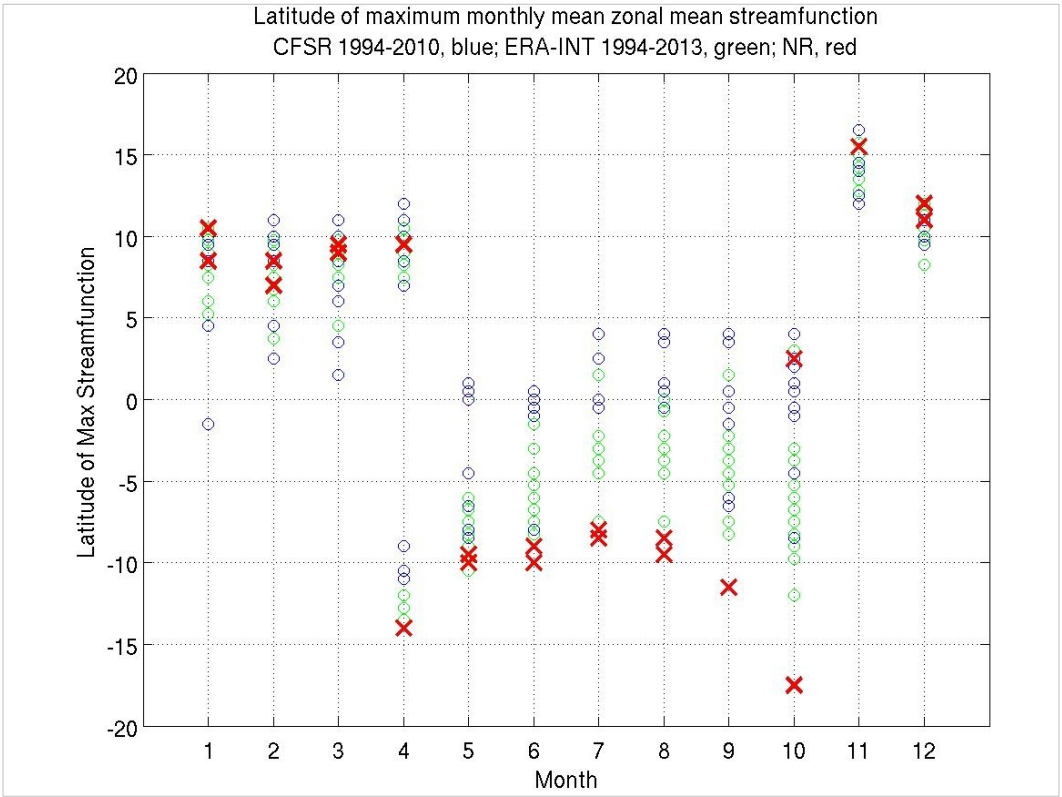
Comparison of temporal and spatial Interpolation errors compared to 1.5 km run for Typhoon Guchol (2012).

It is common with large global NR to save the full spatial field at intervals of 30 minutes-3 hrs. To make the simulated observations, the spatial fields are interpolated with bilinear methods, and linear temporal interpolation. However, time interpolation can result in not just a loss of high spatial resolution features (which often have short timescales), but deformation of large scale features that evolve in time. Thus, the interpolated field being sampled to generate synthetic observations may be much less realistic than the model output. It is best to use as frequent temporal output as you can afford.

Nature Run Validation

- Evaluate if NR is sufficiently realistic to yield meaningful results
- In addition to the phenomena of interest, the NR needs to realistically replicate fields needed to generate synthetic observations
- Can't validate everything; corollary – don't expect a NR to come pre-validated for your needs

It is best to perform at least some sanity check validation of the NR for your given experiment. Specifically, validate for any fields needed to generate your new observation type, and also factors that influence the metrics of interest for your OSSE.



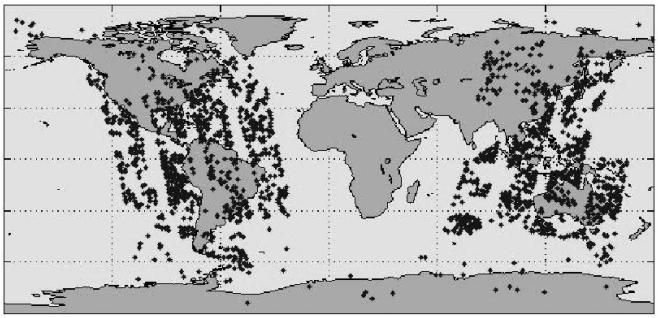
When validating the NR, one would like any section of the NR to be indistinguishable from a randomly drawn period from the real world. In this figure is an example of validation of the G5NR (GMAO NR). Monthly mean streamfunction statistics (details unimportant for this illustration) are compared for individual months of the NR (red X) and 15-20 years of months from two reanalyses (green and blue circles). Ideally, the NR values should fall within the envelope of the real world examples.

Here we see that for some months (Feb, March, Jan, Nov, Dec), this is true. However, the NR values are extrema or outside of the reanalysis envelope during the boreal summer.

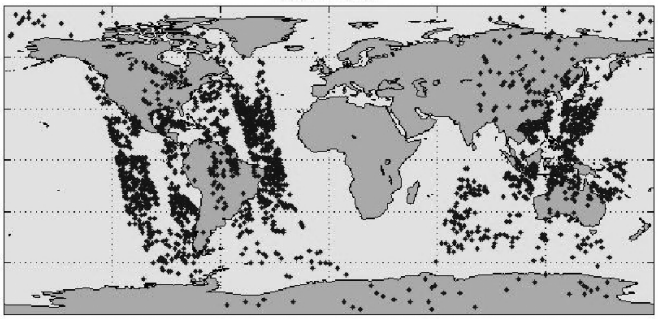
Synthetic Observations

Example of AIRS observations channel 295 at 18 UTC 12 July

Simulated



Real



This is for a channel that peaks in the lower troposphere and is thus easily affected by clouds. The blank patches in the observing swaths result from QC rejections due to the presence of clouds. The simulated and real results should not be identical because, at any particular time, the real and NR clouds may be in different locations. The character of the plots should be similar however, like the typical sizes of the cloudy patches, wherever they may be. In this example, the counts of obs are almost identical

The most common practice for making synthetic obs for existing data types is to use the archived temporal and spatial distribution of observations as a basis for the synthetic observations. The observation value itself is interpolated and/or calculated with an observation operator from the NR fields, using the time and location of corresponding real observations as a base. This works for most data types, but needs to be modified to account for factors that are influenced by the atmospheric fields themselves – ie cloud contamination of radiance observations, or the spatial and temporal location of AMVs (which depend on the NR cloud field), or the advection of raobs/dropsondes by the NR wind field as they ascend/descend.

Observation Errors

- Synthetic observations contain some intrinsic interpolation/operator errors, but less than real observations (usually)
- Synthetic errors are created and added to the synthetic observations to compensate
- Error is complex and poorly understood
  - Error magnitude
  - Biases
  - Correlated errors

Intrinsic observation errors in the OSSE stem from differences in model resolution between the NR and the DAS (and interpolation as previously shown) and due to flaws in the method by which the observation is simulated. It is preferred that a different operator is used to generate the observations compared to those used to ingest the observations (ie, different radiative transfer code, different methods of GPS, etc). Regardless, in practice these intrinsic errors are smaller than the errors associated with real observations, at least as far as we can judge.

Synthetic errors can have several components: random uncorrelated errors are the easiest and most commonly used type of synthetic error, however the DAS is very skilled at removing random uncorrelated errors when ingesting obs. What the DAS cannot easily remove are correlated errors, which find their way into the analysis, affecting the analysis increment, analysis skill, and forecast skill much more than random errors. These error correlations are very difficult to identify in the real world and cannot be disentangled from model error.

Because OSSEs tend to have insufficient model error, when matching statistics of the DAS behavior by manipulating observation error it is entirely possible to end up overcompensating for the lack of model error by overinflating the correlated observation errors. This is undesirable but can be difficult to identify and avoid.

Biases that we understand are removed by bias correction; biases that we do not understand are unknown. If we add the known biases, these will be immediately removed by the bias correction procedures of the DAS. The bias correction of the DAS should be adjusted to match the synthetic biases as the spinup time for bias correction can be long.

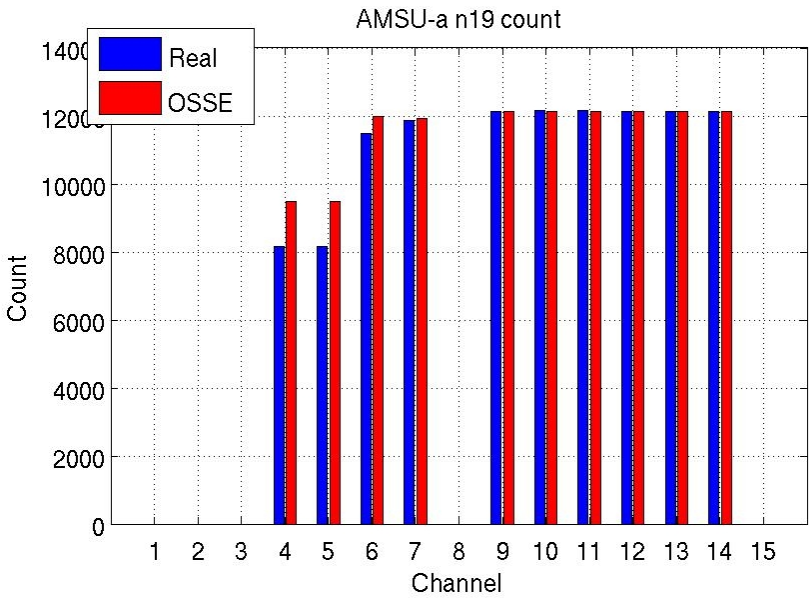
Calibration

- Adjust synthetic observations and their errors to increase realism of the OSSE in a statistical sense
  - Compare OSSE statistics to statistics using real data in the same DAS/forecast system
- Need to decide what statistical metrics to use for the calibration, depending on your needs
- Calibrating new observation types?
  - Find an analogous data type if possible

Calibration is the process by which we adjust the OSSE to match the real world in the ways that we care about most (if possible) and demonstrate how well the results of the OSSE can be trusted. To do this, we run a set of forecasts using the desired experimental model with real data for a sufficient time period (depends on the statistics you are interested in), and then run a complementary test using the synthetic observations.

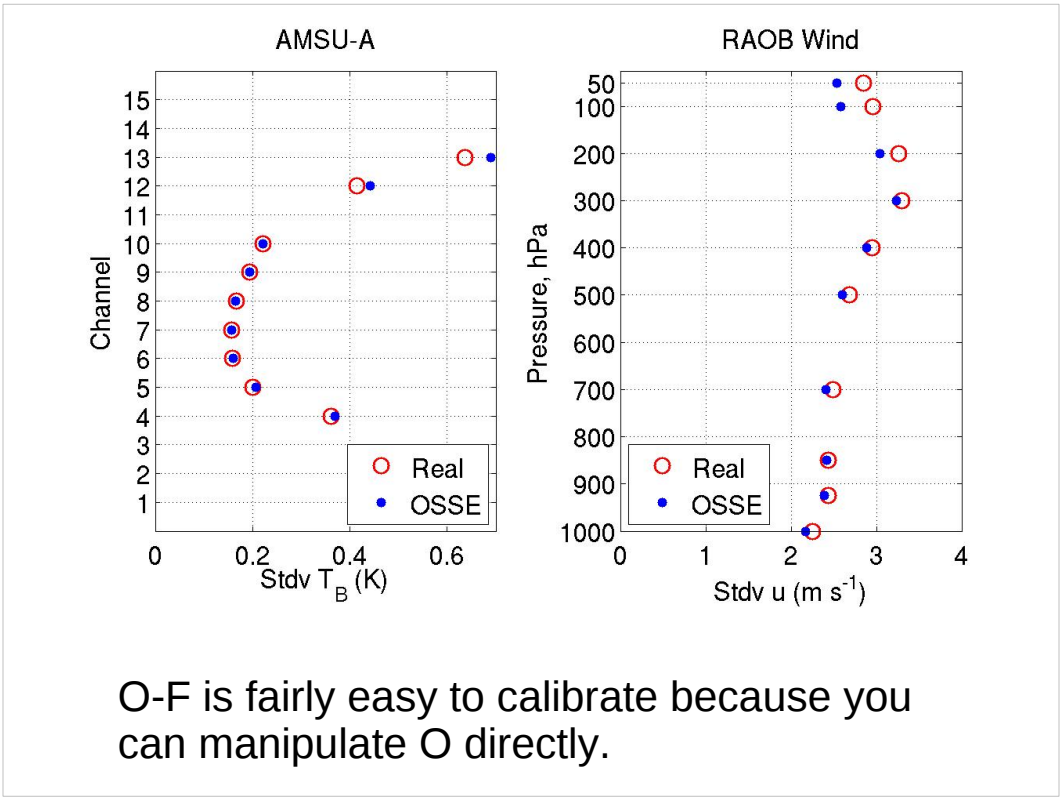
Ideally, we would be able to match up all statistics of the DAS and forecast skill between the OSSE and the real world, but in practice we can only hope to match up a few. It is important to choose beforehand which are the most important metrics that you want to match depending on the goals of your OSSE.

Calibration of new observing types is difficult because there is no real world data for comparison. If there is a similar observation type, this could be used as the basis of calibration. Otherwise, you may wish to test the robustness of your results by repeating experiments with different magnitudes and types of errors on the new observations.

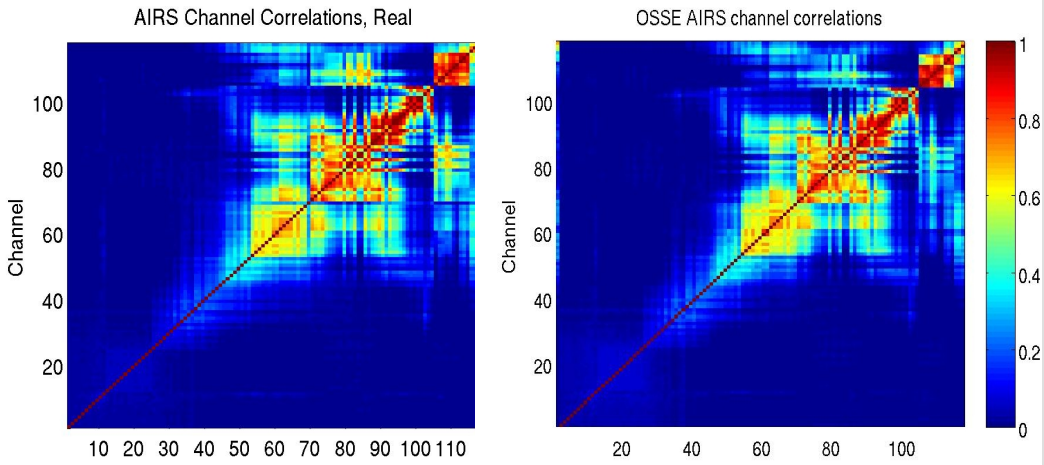


Observation count is easy to calibrate

Observation count – this can be manipulated not just by changing the observation error, but also by the characteristics of the observations themselves (thinning, locations, etc). The goal is to match the count of observations that survive the quality control process and are ingested by the DAS, which can require manipulation of the synthetic observations to model the influence of gross errors.

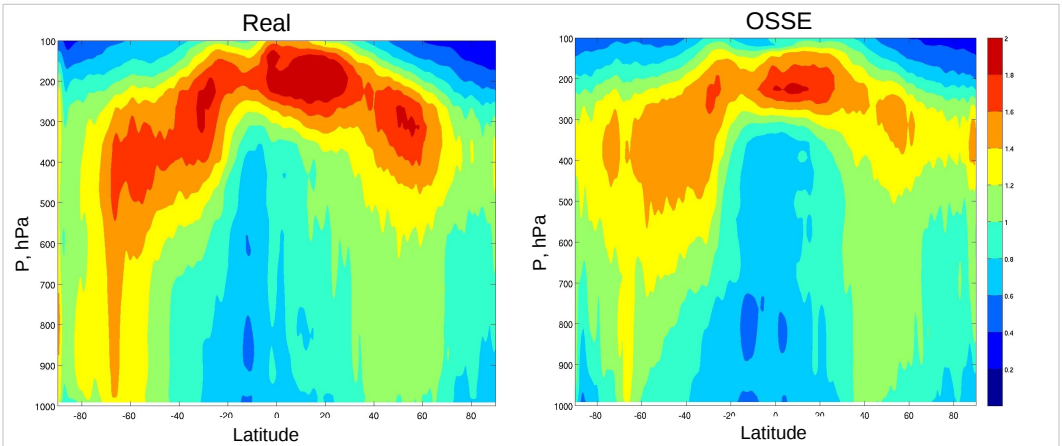


It turns out that F (the background) is relatively stable when manipulating the observation errors. It may only take 2-3 iterations of error adjustments on the observations to closely match O-F variances between the OSSE and the real world. This can still be difficult in cases where observation counts are low (see upper tropospheric/stratospheric RAOBs above) or where the NR and experimental model have differences of opinion about climatology. Observations near the surface can also be difficult to calibrate.



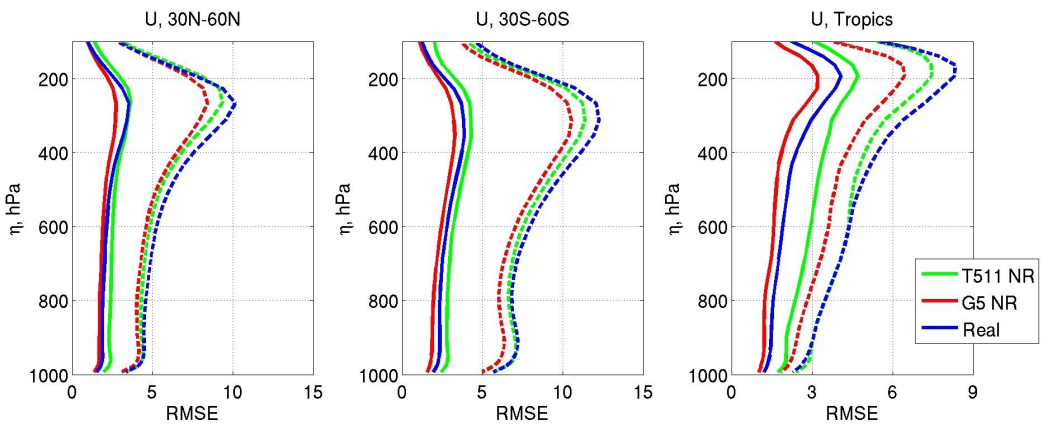
Some observation correlations are relatively easy to calibrate

These channel correlations can be calculated directly for the real satellite channels, and are generally well-replicated in the synthetic observations (after adjustment of correlated errors). However, other types of correlations, such as horizontal and vertical error correlations can be much more difficult to replicate as they are not well-understood in the real world. Not just the magnitude, but the spatial (and temporal) character of the error correlations should ideally be reproduced in the OSSE, but it is not clear how to accomplish this is practice.



A-B (analysis increment) is a little harder to calibrate, as A and B are not directly controlled

Analysis increments are not directly controlled by observation errors in the way that O-F is. In this case, in order to match the A-B, the observation errors would have to be inflated such that O-F no longer match. The discrepancy between the OSSE and the real world in this case is suspected to be largely due to insufficient model error, and possibly in part due to incorrect correlation of observation errors: A-B is balanced when cycling exactly by error growth during the 6-hour period between cycles. If that error growth is too weak, A-B will also be too small.



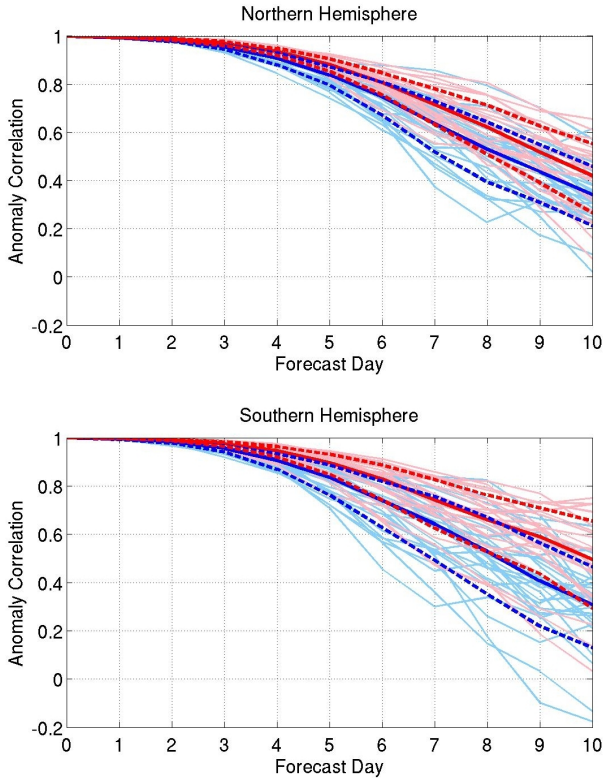
Forecast errors are harder to calibrate, especially for longer forecasts. Matching of this statistic by manipulation of observations is difficult to impossible beyond ~24 hour forecasts.

In order to match these forecast skills, the synthetic observations would have to be clobbered by observation errors. While the forecast skill might then “match” the real data case, the observation impacts would likely be devastated, and the point of performing the OSSE would be lost.

Model error determines forecast skill in the longer term forecast, so calibration is not possible (unless you want to mess with your model).

Red: OSSE
Blue: Real

500 hPa anomaly correlations of geopotential height

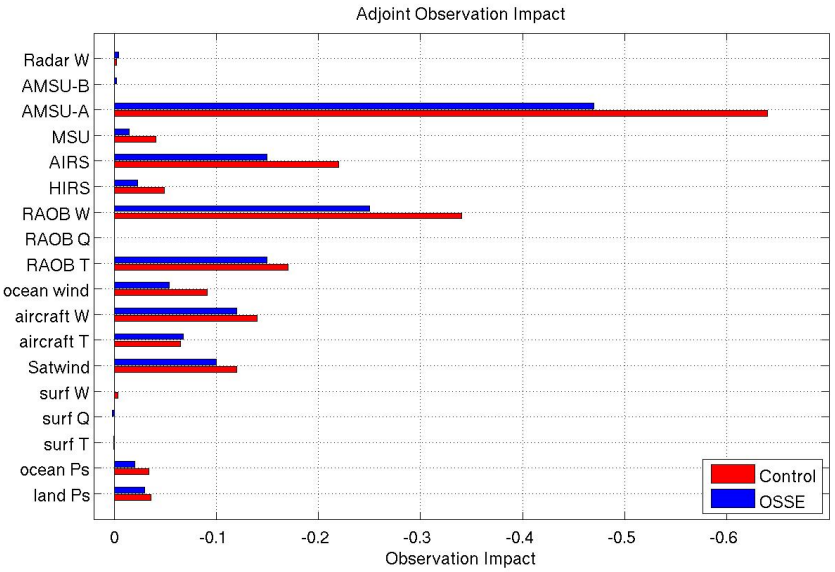


The OSSE in this example has significantly higher anomaly correlations than in the real world, even out to the day 10 forecast. The OSSE also has fewer dropouts (very low scoring forecasts). Even so, this OSSE can still be used for experiments, as long as the too-high skills are accounted for.

"Accounting for" deficiencies in the OSSE framework is often mentioned but rarely discussed (or practiced). At the very minimum, deficiencies should be explicitly described in any published work, and in communication with decision-makers who may place weight on results from the OSSE.

If the OSSE has flaws which can significantly impact the results, it can be difficult to estimate how these flaws might actually affect the observation impacts. In the above example, the model error is too small, resulting in excessive forecast skill. On the one hand, a new observing system might have reduced impact in the OSSE compared to the real world, as the background state is "too good", so there is less "work" to be done by the observations during DAS. On the other hand, any information introduced by the new observations may be retained longer into the forecast integration compared to the real world, where model error might act to quickly erase the added information. These effects depend on the character of the information added by the observation, the dynamics of how this information evolves in time, and the nature of model error in the OSSE compared to the real world – ie, a very complex and nonlinear interaction.

Why believe OSSE results?



New observations can be put into context relative to existing observation impacts

Given that calibration cannot fix all the problems with forecast skill, analysis and background quality, etc, what can we actually get that is useful from the OSSE? One metric that can frequently yield good results is observation impact – whether by data denial experiments or adjoint (above bar graph).

Observation impacts can be manipulated somewhat by adjusting observation errors – however the interplay between observation errors and impact is complex.

For data denial experiments, increasing the observation errors increases the magnitude of the effect the observations have on the analysis state (increasing the analysis increment) up to the point where quality control begins to eliminate the observations. In this case, the effect of the observations is not judged as to whether it has a positive or negative impact on the analysis state.

In the adjoint, where observations are judged by their impact on the 24 hour forecast skill, increased observation errors have complicated effects. If the observation errors are ingested and retained in the analysis, these errors may grow and increase the error of the background state of the next analysis state, thus allowing the observations to do more work (and have a larger impact) at the next cycle time. If the observation errors are increased on all observation types, many of the data types may see (good) increases in the adjoint impact. If only one data type has increased errors, however, the impact of that data type may be reduced while the impact of other data types increases to compensate for the greater error of the background.

In general, adjoint impacts in OSSEs may be smaller for all observing types than for real data, due to the too-skillful background state from insufficient model error. However, the relative impact of different data types is well-reproduced, so a new observing type can be put into context with the current observing systems.

Criticisms of OSSEs

- Results only apply within the OSSE system – no concrete connection to the real world
- Even the best OSSEs are far from perfect: incestuousness, difficulty in generating observations and errors, deficiencies of the Nature Run
- By the time the new instrument is deployed, both the global observing network and the forecast models/DAS will be different
- Examples of sloppy or unsuccessful OSSEs

These are all true! We need to recognize the limitations of the OSSE results. There are however, few alternatives and we should not let the perfect be the enemy of the good.

Take care when reading the literature. Did the investigators validate their NR? Did they calibrate their OSSE? Was the new observing type tested in an honest comparison against the full suite (or a reasonable suite) of existing observation types? Were the results cherry-picked from a case study or tested more robustly?

Common Pitfalls

- Very reduced baseline of assimilated observational data (ex. no radiance data)
- Other artificial degradation of analysis state
- No validation or calibration of OSSE framework
- Obtaining robust results from case studies is very challenging
  - Use ensemble forecasts if you can!

Because of the expense of generating synthetic observations, one shortcut is to omit the more difficult (usually radiance) observations from the synthetic obs dataset. This is especially common in regional OSSE studies. The problem is that conventional observations can be quite sparse, and the comparison is neither fair nor realistic when judging a new observation type. If you just want to make sure a new data type is functioning properly, then this type of limited study can be successful, but it is not appropriate when trying to estimate the potential impact of a new observation in operational context.

When a full-blown OSSE is not performed – ie, analyses are used instead of a NR or resources are otherwise limited, the investigator can have a problem of how to make a background starting point for the DAS which is different from the truth. Sometimes the investigators will chose to damage the analysis state in some way to make a new background state. This is undesirable, as the background state is influenced by the entire observational suite, including the new observations, so the "work" done by the observations becomes completely artificial and dependent on how the background was degraded rather than an honest test of the observation impact.

Unvalidated/uncalibrated systems yield results which cannot be trusted or put into context in reality. What do you learn in this situation?

Sometimes case studies are the only option available. Interpreting the results can be very challenging – the use of an ensemble of forecasts can be much more robust than a single deterministic forecast.

Choosing Metrics

- Long cycling periods necessary to get statistically significant results for most new observations
- Anomaly correlation is a difficult metric to show appreciable impacts
- What fields do you expect the instrument to improve?
- Largest impacts found at analysis time or short-term forecasts

Most new added data types will have only a tiny impact on forecast skill, so if you want statistically significant results, you will need to run tests over a lengthy period to gather enough data. One strength is the use of paired statistics, which can make it easier to show significance with a smaller dataset (see, for example, Wilks' “Statistical Methods in the Atmospheric Sciences”).

Some metrics are especially hard to move – anomaly correlation is one of them. Instrument teams and administration also have a hard time with anomaly correlation as a concept (and as “great news – your instrument increased 5-day anomaly correlations by 0.01!!”). Instead, consider other metrics that more directly reflect the impact you expect from the new instrument.

Idealized Studies

- Identical twin experiments
- Idealized observations
- Manipulation of observation errors
- Experiments with **B**, **R**

- Make use of available “Truth”

OSSEs can also be used as a platform for experimenting directly with DAS processes and behavior. You know the true state of the atmosphere, have direct control over the observations, a large degree of control over the observation errors, and can manipulate the observations and characteristics of the DAS. At the same time, the system is much more complex than a simple toy model.

Regional OSSEs

- Regional OSSEs are harder than global OSSEs
  - Two Nature Runs (local embedded in global)
  - Two forecast models each using synthetic obs
  - Shortcuts are very often taken – need to carefully examine the methods to ascertain if the results are trustworthy

Embedding a regional NR inside of a global NR is hard, especially if you want the behavior in the regional NR to match the large scale behavior in the same region of the global NR. See recent work at HRD (Nolan et al 2013, J. Adv. Earth. Model. Syst) for an example of a true regional NR.

In practice, regional OSSEs often use a forecast or analysis from a global model (such as the GFS or ECMWF) and use that BOTH for the global NR and for the global forecast model (ie for the boundary conditions fed to the regional forecast experiment model and the regional NR). Care should be taken that the influence of the boundary conditions does not override the influence of the initial conditions/obs impact in the region model during the forecast experiment. For new observing systems that would act on a wide area, this particularly neglects the influence of the new obs on the global model then improving the boundary conditions of the regional forecast model and thus the entire field.

Takeaway

- OSSEs can provide useful information about new observational types and the workings of data assimilation systems
- Careful consideration of research goals should guide each step of the OSSE process
- OSSEs are hard, good OSSEs are harder

References

Review article: 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. Environ., **115**, 199-213.

Observation errors: 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

Regional OSSEs: Nolan, David S., Robert Atlas, Kieran T. Bhatia, and Lisa R. Bucci, 2013: Development and validation of a hurricane nature run using the Joint OSSE Nature Run and the WRF model. J. Adv. Earth. Model. Syst., **5** , 1-24

Early history of OSSEs: Arnold, C.P. And C.H. Dey, 1986: Observing System Simulation Experiments: past, present, and future. Bull. Amer. Meteor. Soc., **67**, 687-695.