Evaluation of adjoint-based observation impacts as a function of forecast length using an Observing System Simulation Experiment

N.C. Privé*,[a,b] Ronald M. Errico[a,b] and Ricardo Todling[b] and Amal El Akkraoui[b,c]

[a]*GESTAR, Morgan State University, Code 610.1 NASA/GMAO, Greenbelt, MD, USA 20771*

[b]*Global Modeling and Assimilation Office, Code 610.1 NASA/GMAO, Greenbelt, MD, USA*

[c]*Science Systems and Applications, Inc., Greenbelt, MD, USA*

*Correspondence to: GESTAR, Morgan State University, Code 610.1 NASA/GMAO, Greenbelt, MD, USA 20771. E-mail:nikki.prive@nasa.gov

Adjoints of numerical weather prediction models may be employed for Forecast Sensitivity to Observation (FSO) in order to monitor the contribution of ingested observation data on short-term forecast skill. However, the calculation of short-term forecast error is difficult due to the lack of a truly independent dataset for verification. In an Observing System Simulation Experiment framework, the Nature Run is able to provide a true and complete verification dataset and allows accurate evaluation of short term forecast errors. In this work, an OSSE developed at the National Aeronautics and Space Administration Global Modeling and Assimilation Office is used to explore the impact of observational data on forecasts in the 6 to 48 hour range. An adjoint of the Global Earth Observing System model is employed to compare the observation impacts estimated using both self-analysis verification and the true Nature Run verification.

Self-analysis verification is found to inflate the estimated forecast error growth during the early forecast period, resulting in overestimations of observation impacts, particularly in the 6-12 hour forecast range. By 48 hours, the self-analysis verification estimates of forecast error and observation impacts more closely match the true values. The fraction of beneficial observations is also overinflated at short forecast times when self-analysis verification is used. The progression of impacts of an individual observation or data type depends on the character of the growth of the initial condition error that each observation affects.

Key Words: numerical weather prediction, forecast sensitivity to observations, adjoint models, OSSEs

Received . . .

1. Introduction

Millions of observations of the atmosphere are ingested by data assimilation systems (DAS) every day as a crucial element of operational numerical weather prediction (NWP). The contributions of these observations to the forecast skill can be monitored and assessed by employing one of the many variations of what is referred to as forecast sensitivity to observations (FSO).

One method of FSO uses adjoint models in order to calculate the impact that all ingested observations have on a selected error norm without the need to run multiple data denial experiments (Baker and Daley 2000, Gelaro and Zhu 2009). The Trémolet (2008) extension of the Langland and Baker (2004) approach to FSO uses pairs of forecasts in order to estimate the observation impact on forecast skill. One forecast starts from an analysis field, while the second forecast can be thought of as starting from the corresponding background field at the same analysis time. The difference in the initial states of these two forecasts is due solely to the ingestion of information from observations via the DAS. As each of the two forecasts integrates forward in time, any differences in forecast skill are considered to be the result of the injection of information by observations at the initial analysis time.

The impact of a particular data type or individual observation is a function of the forecast length, as the error in the background field that is corrected by the observation(s) grows and/or decays with model integration in time. Background errors may project onto structures that peak at the initial time and then rapidly decay, structures that grow exponentially before saturation, or structures that are overtaken by model error growth, among many possibilities. Some of these differences in error structures may vary regionally, such as the difference in error growth in the tropics where there are fast convective processes and substantial model error as compared to the extratropics, where baroclinic dynamics may have errors that grow with longer timescales.

Because adjoint models employ a linearization of the forecast model, relatively short-range 24-hour forecasts are often selected when using FSO. However, short forecasts present a challenge in terms of verifying the forecast error for adjoint calculations, particularly when the self-analysis field is selected to serve as the 'true' state of the atmosphere.

Some recent studies have examined the influence of the choice of verification on estimates of FSO. Necker *et al.* (2018) compared the use of a set of independent radar observations versus subsets of ingested observations for verification with ensemble FSO. They found that biases in the verification fields had strong effects on the estimated observation impacts. Kotsuki *et al.* (2019) looked at verification methods with ensemble FSO for short forecasts of 6 to 12 hours, comparing self-analysis verification to verification with reanalysis and observations. In their study, using self-analysis verification resulted in overinflated fractions of beneficial observations, particularly at 6 hours.

The effects of verification on FSO with adjoint models have also been explored in several studies. Daescu (2009) showed the mathematical basis by which uncertainty in the verification field could result in uncertainty in the calculations of observation impacts. A general expression for the error in self-analysis verification is given in Todling (2013) for any length of forecast. Cardinali (2018) used observations as verification and compared the results with self-analysis for 24-hour forecasts. Jung *et al.* (2013) found high fractions of beneficial observations at 6 hours using self-analysis verification.

Observing system simulation experiment (OSSE) frameworks can be very useful for investigating the behavior of data assimilation systems and the evolution of short-term forecast skill. In an OSSE, the real world is replaced with a simulation from a high resolution NWP model; this simulation is called the Nature Run (NR) and is considered to be the 'truth'. Observational data are simulated using the NR fields for the same data types used in operational NWP, and are ingested into a different NWP model. Because the 'truth' is completely known in the form of the NR, the short term forecast error can be explicitly calculated. Kotsuki *et al.* (2019) suggested the use of an OSSE to determine the cause of exaggeration of observation impacts with self-analysis verification.

Such an OSSE system has been developed at the National Aeronautics and Space Administration Global Modeling and Assimilation Office (NASA/GMAO; Errico *et al.* 2017). The GMAO OSSE framework includes different versions of the Global

Earth Observing System Model (GEOS; Rienecker *et al.* 2008) used to make the NR and the experimental forecasts, as well as the Gridpoint Statistical Interpolation (GSI; Kleist *et al.* (2009)) data assimilation system. This OSSE framework has been extensively validated to ensure that the performance is robust and gives meaningful results (Errico and Privé 2018, Errico *et al.* 2013).

An adjoint of the GEOS model is available (Holdaway *et al.* 2014) and can be used in the OSSE framework to explore the behavior of observation impacts at short forecast times. The first aspect of the FSO that is of interest is the evolution of observation impact from the 6-hour to the 48-hour forecast. The progression of observation impacts on forecasts of increasing length can be characterized for various data types and regions. The adjoint also allows the evolution of the impacts of individual observations to be traced.

The second aspect of the FSO that will be explored is a comparison of the observation impact estimates calculated with self-analysis verification versus with the 'true' NR verification. This is of particular interest as the NR verification is not available outside of the OSSE framework. While observation impact estimates are expected to have better accuracy for short-range forecasts than for long-range forecasts due to linearization limitations, self-analysis verification introduces undesirable correlations that are larger at short ranges than at longer ranges in the forecast. By comparing the two verification methods in the OSSE context, the range of forecasts for which the adjoint gives useful results with self-analysis verification can be estimated.

Details of the OSSE framework used in these experiments and of the adjoint operator are described in Section 2. The evolution of observation impacts at different forecast lengths is explored in Section 3, and the comparison of verification methods in Section 4. Some overall conclusions are discussed in Section 5.

2. Method

A numerical weather prediction OSSE framework has been developed at the National Aeronautics and Space Administration Global Modeling and Assimilation Office (NASA/GMAO), and is used for all experiments here. In addition to the standard validation techniques (Errico *et al.* 2013, Privé *et al.* 2013b), validation of the adjoint tool and early forecast error has been performed and is described in Section 2.2.

2.1. *Experiment Framework*

The GMAO OSSE framework uses a Nature Run developed in-house and commonly referred to as the "G5NR" (Gelaro *et al.* 2014). The G5NR is a free run of the 2012 version of the Global Earth Observing System Model, at approximately 7-km horizontal resolution with 72 vertical levels, for a two year integration. The G5NR uses archived boundary conditions for sea surface temperatures and sea ice from the 2005-2007 time period, and thus has date-stamps that refer to this time range. However, there is no expectation of synoptic agreement between these dates in the G5NR and the same dates in the real world.

Simulated or "synthetic" observations are generated for most of the data types that were operationally ingested at NASA/GMAO in 2015. These simulated observations are meant to mimic real observations. For some conventional data types such as surface observations and aircraft observations, the locations and times of real observations from 2015 are used to interpolate the G5NR fields at the same spatiotemporal locations to create the synthetic data. For rawinsondes, the launch times are taken from real data archives but the rawinsondes drift using the G5NR wind fields. Atmospheric motion vectors are treated differently than other data types, with the synthetic data completely dependent on the distributions of clouds and water vapor in the G5NR for congruity (Errico *et al.* (2020)).

Radiance data including AMSU-A, AIRS, HIRS-4, SSMIS, IASI, CrIS, and MHS are generated using the locations and times of real data, employing the Community Radiance Transfer Model (CRTM; Han *et al.* 2006) with the G5NR fields to generate the synthetic observations. These simulated radiance observations are affected by the G5NR cloud field to produce observation locations so that the selection of cloud free observations by the DAS is consistent with the NR synoptic state. GPS-RO data are created using real locations of GPS-RO, using the G5NR fields with the Radio Occultation Meteorology Satellite Application Facility software (Culverwell *et al.* 2015). Full details of the observation simulation process are described in Errico *et al.* (2017).

The synthetic observations when generated do not have the same error characteristics as real observations. Simulated errors are added to the synthetic observations to match certain statistical characteristics of real data. For example, during the calibration process, the statistics of observation counts ingested into the data assimilation system (DAS) and the variances of observation innovations are matched as closely as possible between the synthetic data and real data. Additionally, the magnitude of correlated and uncorrelated errors added to the synthetic observations are adjusted in an iterative process until these statistics are as close as possible to those of real data. Uncorrelated random errors are added to all synthetic data types; horizontally correlated errors are added to AMVs, AMSU-A, HIRS-4, SSMIS, and MHS; channel-correlated errors are added to AIRS, IASI, and CrIS; and vertically correlated errors are added to rawinsonde, AMV, and GPS-RO observations. Biases are not added to the synthetic observations, as the only biases that are understood are those that are removed by the bias correction scheme used in the DAS. However, the GSI bias correction routines are allowed to act upon the radiance data, with bias coefficients that were spun up for several weeks of the OSSE assimilation prior to the start of the experiments.

The synthetic observations are ingested by the GSI in its three-dimensional variational data assimilation form using the First Guess at Appropriate Time approach (FGAT; Lawless 2010 and Massart *et al.* 2010). The GEOS model version 5.17 at C360 resolution on the cube-sphere (approximately 25 km horizontal resolution) is employed for forecasts. This version of the GEOS model is approximately five years more recent than that used to generate the G5NR, and includes some substantial differences in model physics, including the switch from single moment to two moment microphysics (Barahona *et al.* 2014). These changes result in some model bias between the G5NR and the forecast model, but with less model error than would be expected in the real world. This framework can be considered a "fraternal twin" OSSE.

The OSSE model run and data assimilation begin on 10 June 2006 in the NR timeline, with a spinup period of 20 days. The OSSE is cycled through 31 August 2006, treating the period of 1 July to 31 August as the experimental timeframe.

The GEOS adjoint model has a moist component that accounts for convective processes (Holdaway *et al.* 2014). For all FSO calculations in these experiments, the total wet energy (e) norm is used (Ehrendorfer and Errico 1995), as defined by

$$\begin{aligned} e \;=\; & \frac{1}{A}\sum_{i,j,k}\frac{1}{2}\Big[u_{i,j,k}^{\prime 2}+v_{i,j,k}^{\prime 2}+\frac{c_p}{T_0}T_{i,j,k}^{\prime 2}+ \\ & RT\Big(\frac{p_{s,i,j}}{p_0}\Big)^{\prime 2}+\epsilon\frac{L^2}{c_pT_0}q_{i,j,k}^{\prime 2}\Big]\delta A\delta\sigma_k \qquad (1) \end{aligned}$$

where $u^{'}$ and $v^{'}$ are the zonal and meridional wind errors, $T^{'}$ is the temperature error, $q^{'}$ is the specific humidity error, A is the area and σ_k is the fractional mass in the kth model layer for the column of air at the i, jth horizontal gridpoint, L is the latent heat of condensation, c_p is the constant specific heat capacity of air, $T_0 = 270.0$K and $p_0 = 1000.0$hPa, R is the gas constant of dry air, and ϵ is an assigned weighting of the humidity term, here chosen to be 0.3. This norm is calculated for the layers between the surface and 0.7 hPa.

The FSO experiments explored in this work involve energy norms calculated for different forecast lengths. A single run of the OSSE and cycling DAS is used throughout the comparisons that follow, with pairs of forecasts initiated at 1800 and 0000 UTC each day. FSO is calculated for 6, 12, 24, and 48-hour forecasts. In each case, two sets of FSO results are obtained, one by verifying the corresponding forecasts with the NR fields (Section 3), and another by self-verifying (Section 4) as is typically done in real operational NWP settings.

2.2. *Validation*

Validation is important when working in an OSSE framework, considering that all aspects of the OSSE are simulated, but we use the results of that simulation to infer what occurs in reality. In these experiments, validation of the adjoint estimates of observation impact is critical, as is validation of the analysis error and forecast error growth, since the observation impacts are the primary metric of interest. The real data case used for validation employs the same GEOS model version, starting on 11 June 2015 with 20 days of spinup, and validation period of 1 July 2015 to 31 August 2015. This time period is chosen to coincide with the period used as the basis for the generation of synthetic observations for the OSSE. Although the synoptics of the real data case differ from those in the OSSE, the global observing network is as similar as possible.

Figure 1 shows the daily mean adjoint estimate of observation impact on 24-hour forecast skill with self-analysis verification for the real data case and OSSE case. For most data types, the estimated observation impact is considerably smaller (40-60%) for the OSSE than for the real data, with the exception of rawinsonde humidity and AMVs. This result of smaller impact is common to NWP OSSEs (Privé *et al.* 2013b), and generally thought to be caused by insufficient model error in the OSSE. While there are differences between the model version used for the NR and that used for the forecasts, there is less model bias and smaller variance of model error than is expected in the real atmosphere. Lack of model bias could contribute significantly to the smaller magnitudes of observation impacts seen in the OSSE, and will be discussed further in Section 5. However, the overall relative ranking of observation types by impact in the OSSE is similar to real data.

The analysis increment is selected to validate the amount of "work" done by the observations during data assimilation. The zonal mean root temporal mean square (RMS) of the analysis increments (A-B, where A is the analysis state and B is the prior background state) for temperature and zonal wind are shown for the Real and OSSE cases in Figure 2. The RMS of the analysis increments are approximately 30% lower in the OSSE as compared to Real. The spatial structure of the analysis increment is similar in both cases. This agrees with the adjoint estimates of observation impact having smaller magnitude in the OSSE. These results imply that there is insufficient forecast error growth in the OSSE, as the magnitude of the analysis increment should balance the growth of errors between cycle times (6 hours) if the statistics of the analysis error are generally stable in time.

Note that it would be possible to increase somewhat the error in the OSSE during the initial forecast period by adding correlated errors with greater magnitude to the synthetic observations. However, this would cause the temporal variance of observation minus background to be greater in the OSSE as compared to real data, and would likely be artificially compensating for insufficient model error, at least in part. The magnitude of the errors needed to alter the adjoint impact estimates would actually be quite large (Privé *et al.* 2013a). Instead, we have preferred here to match the observation innovation statistics while keeping in mind that the OSSE adjoint shows smaller impacts when interpreting the results.

The short term forecast error growth can be used to inform expectations of the OSSE performance for adjoint estimates of observation impact on forecast skill. Figure 3 shows the short term global root mean square error (RMSE) for temperature (Fig. 3a) at 506 hPa and zonal wind (Fig. 3b) at 226 hPa over the 48 hour forecast period (these are internal model η levels). Three sets of RMSEs are shown: the Real case using self-analysis verification (heavy solid line); the OSSE case using self-analysis verification (thin solid line); and the OSSE case using the NR as verification (dashed line), i.e. the true error. As expected, the self-analysis verification forecast error for the OSSE severely underestimates the true forecast error at short forecast times but approaches the NR-verified error at longer forecast times. The self-analysis verified forecast error in the OSSE is approximately 20-25% lower than the forecast error for the Real case. However, the functional form of the RMS forecast error growth in the OSSE case is similar to that in the Real case. While there are substantial differences between the Real and OSSE case, the consistency of these differences over the range of forecast lengths is encouraging that the OSSE adjoint results are applicable to the real world with suitable adjustments to the magnitudes of observation impact.

3. Evolution of Adjoint Impacts

The adjoint tool relies on a linearization of the forward numerical weather prediction model to estimate the evolution of perturbations. This linearization is expected to diverge from the behavior of the full forward model as the forecast time increases. The observation impact totalled for all data types captured by the adjoint tool at each forecast length (open circles) is compared to the nonlinear net impact as the solid black circles in Figure 4a. This nonlinear net impact is the difference in error between the pairs of forecasts initialized six hours apart. The magnitude of the nonlinear impact increases nearly linearly with forecast length over the first 48 hours, where negative impacts indicate a decrease in forecast errors due to the ingestion of observational information. The adjoint estimate of observation impact also increases in magnitude with forecast length, but with a slower rate of increase and smaller magnitude overall. The fraction of the nonlinear impact captured by the adjoint tool (squares in Figure 4b) decreases from approximately 90% at the 6 hour forecast to 64% at 48 hours when NR-verified.

As the magnitude of the observation impact grows over the first two days of the forecast length, it is expected that the net observation impact must eventually decrease and approach zero. This is because the forecast error will asymptote toward a steady magnitude as all forecast skill is lost and errors saturate, generally sometime after the two week forecast length. As long as the forecast remains bounded by a realistic climatology, the RMS forecast error will be bounded by the RMS difference between pairs of randomly chosen synoptic states. A schematic of the growth and saturation of this type of error is illustrated by the solid line in Figure 5a, with the corresponding observation impact in 5b. This observation impact behavior is expected to occur for initial condition errors that project onto growing structures that see peak growth after the initial forecast period and then decay or reach a saturated state. However, the majority of initial condition errors project onto structures that decay, remain constant, or are swamped by model errors during the early forecast period (Errico *et al.* 2001). The observation impacts that are associated with these fast timescale error structures will therefore have the greatest magnitude at the initial forecast time and decrease in magnitude as the forecast progresses (dashed line in Figure 5). However, due to the linear nature of the adjoint model, the adjoint estimate is expected to grow unbounded as time increases (Legras and Vautard 1996).

The normalized adjoint estimates of observation impact for each of the different forecast lengths (6, 12, 24, and 48 hours) are shown in Figure 6 for three regions: the northern hemisphere extratropics from 70°N to 20°N (NHEX), the southern hemisphere extratropics from 70°S to 20°S (SHEX), and the Tropics from 20°N to 20°S. The impacts for each data type are normalized by the 24-hour forecast impact for that type; this normalization is used to make the progression of impacts at different forecast lengths clear for data types having small net impacts. Each observation impact for a data type is made up of thousands or millions of observations over a two-month period, with each individual observation impact potentially projecting onto a multitude of error structures. The net impact behavior of each data type in Figure 6 is a sum of millions of growing and decaying error structures with different magnitudes and timescales, and each line in Figure 3 is a sum of many different lines from Figure 5a.

A variety of observation impacts are displayed by the different data types. The extratropical regions are qualitatively similar in terms of observation impact progression with forecast length for most data types. For global AMSU-A, extratropical ATMS, IASI, SSMIS, rawinsonde temperatures, and AMVs, and NHEX AIRS and aircraft and rawinsonde winds, the observation impact magnitude monotonically increases with forecast length. For MHS, GPS-RO, aircraft temperatures in the extratropics, and aircraft and rawinsonde winds and temperatures in the SHEX region, the observation impacts are nearly constant with forecast length. Rawinsonde humidity impacts in the extratropics diminish in magnitude with increasing forecast length.

The behavior of observation impacts in the Tropics differs substantially from that seen in the extratropics. For most data types, the peak observation impact occurs prior to 48 hours, with some data types having the greatest impact at the 6 hr forecast (AMVs, surface observations, aircraft winds, and rawinsondes). Notably, AMSU-A is the only data type in the Tropical region with monotonically increasing impacts with longer forecast times.

The nature of error growth in the Tropics is expected to differ from that in the extratropics due to the disparate dynamical and physics regimes in these regions. In the Tropics, convective and physical processes with short timescales can lead to rapid growth and then saturation of some types of errors. The humidity field in particular undergoes fast adjustment. Many observation impacts in the Tropics are influenced by processes that are most dominant at the initial time when the model physics act to revert the initial state toward the preferred model climatology or as noisy convective processes that similarly obliterate the information added by observations. For data types that have a peak impact magnitude in the 6-hr to 24-hour forecast length range, the error structures have a short timescale of error growth and saturation, as represented by the dash dot line in Figure 5 with the peak impact close to the initial time.

Error growth in the extratropics is less dominated by the types of short timescale convective and physical processes that are prevalent in the Tropics, and longer timescale error growth associated with large-scale baroclinic and barotropic dynamics plays a greater role. As a result, the progression of observation impacts for some data types follows the solid or dash-dot lines in the schematic Figure 5, with impacts that have peak magnitude for longer forecast times.

In addition to the magnitude of observation impacts, the adjoint tool permits the estimation of the fraction of observations that beneficially (or detrimentally) affect the forecast skill. Past calculations of this quantity by various means (Gelaro *et al.* 2010, Lorenc and Marriott 2014, Hotta *et al.* 2017, and Necker *et al.* 2018) have placed this fraction at slightly higher than 50% for the 24 hour forecast timeframe. Jung *et al.* (2013) and Kotsuki *et al.* (2019) found higher fractions of beneficial observations, as much as 60-70%, for 6 hour forecasts. The expectation is that as the forecast length increases and the error growth reaches saturation, the fraction of beneficial observations will approach 0.50 as any individual observation may be considered to randomly perturb the long-term forecast field. Ehrendorfer (2007) has shown analytically that as observations tend toward uselessness, the fraction of beneficial observations approaches 0.5.

The fraction of beneficial observations as a function of forecast length with NR verification is shown in Figure 7 for the NHEX, SHEX, and Tropics regions. Some data types such as rawinsonde humidities, SSMIS, AMVs and GPS-RO in the Tropics, and MHS in the extratropics demonstrate the anticipated behavior with the largest fraction of beneficial observations at the 6-hour forecast, decreasing toward 0.50 with increasing forecast length. The largest fraction of beneficial impacts are seen for rawinsondes, particularly humidity observations, for forecasts at 6 and 12 hours. These largest fractions are on the order of 55-60%, but decrease to 50-55% by the 24 hour forecast.

Beyond the combined statistics of impacts for particular data types and regions, the adjoint allows the impact of each individual observation to be calculated and traced from the early forecast to the multi-day forecast. A question may be posed as to what the expectation should be for an observation that has a large beneficial (detrimental) impact at a very short forecast time - does this observation impact continue to maintain a large contribution as the forecast integrates forward through the first few days, or could the impact tend toward zero or even switch to being detrimental (beneficial)?

Because of the many types of error growth that may affect the forecast, the influence of observations should be treated statistically. An example of the probabilistic nature of the evolution of impacts of individual observations is illustrated in Figure 8 for AMSU-A NOAA-19 observations in July 2006. The most beneficial and detrimental observations impacting the 6 hr forecast error norm are traced through the 48 hour forecast period. A 2.5σ threshold (where σ is standard deviation) is used to determine which observations occupy the most beneficial and detrimental tails of the distribution of observation impacts. As the forecast progresses, an increasing number of observations switch from beneficial to detrimental, and vice versa. Similar results are found with other data types (not shown).

The mean per-observation impact can be calculated as one method of characterizing the behavior of a select subset of observations. In Figure 9, the evolution of per-observation impacts of several subsets of observations are traced through the lengthening forecast period for forecasts initiated at 0000 UTC for the month of July. The net per-observation impact of all data for several data types (dash dot lines in Figure 9) is slightly negative (beneficial), and remains so as the forecast extends. The distribution of impacts for all observations has a very sharp peak near the mean per-observation impact (not shown). For those observations that are in either the greatly beneficial or detrimental tails of the distribution of observation impacts at the 6 hr forecast time (solid lines in Figure 9), the per-observation impact remains substantial throughout the forecast period, even though the corresponding distributions in Figure 8 show that some observations in the two tails have impacts that change sign at longer forecast times.

Because error growth is often nonlinear, some observations that have the most beneficial or detrimental impacts on the 48 hour forecast may have minimal or even opposite sign impacts at earlier forecast times. The dashed lines in Figure 9 follow the per-observation impact of those observations which occupy the tails of the distribution of impacts for the 48 hour forecast. The progression of impacts for both beneficial and detrimental observations follow an exponential growth pattern, with impacts near zero at short forecast times. Comparing the sets of observations for the tails of the 6 hr and 48 hour impact distributions, approximately 14-27% of the observations that are in the beneficial (detrimental) tail at the 48 hr forecast impact distribution also occupy the beneficial (detrimental) tail of the 6 hour forecast impact distribution. Similarly, approximately 30-40% of the observations with the greatest beneficial impact on the 48 hour forecast skill had detrimental impact on the 6 hour forecast skill. This is a result that should be taken into consideration for approaches that try to selectively eliminate observations deemed detrimental based on a particular measure of impact assessment (Chen and Kalnay 2019).

4. Verification Methods

Figure 10 shows the RMS forecast error as a function of forecast length for both the self-analysis (solid) and NR (dashed) verified calculations. The thin lines are for forecasts starting at 0000 UTC, with the thick lines for forecasts starting at 1800 UTC the prior day, so that the difference beween 1800 UTC and 0000 UTC lines is the impact of the added observations ingested into the 0000 UTC initial time forecast. The forecast RMSE with self-verification approaches the larger magnitude RMSE with NR verification as the forecast length increases. The NR verification RMSE increases nearly linearly with forecast length, while the self-analysis verification RMSE has a greater rate of increase during the initial forecast period. The slope of the forecast RMSE growth is shallower for the NR verification. This indicates that the difference between pairs of 1800 UTC and 0000 UTC forecasts RMSE at any particular verification time is greater for the self-analysis verification calculation than for the NR verification method. These differences between pairs are plotted in Figure 4a, where the total observation impact estimated using self-analysis (solid stars) is 50-75% larger than the NR verification estimate (solid circles) at short forecast times, with the greatest difference at 12 hours.

The adjoint estimation of observation impact (open circles and open stars in Figure 4) is not as strongly affected by the choice of verification as is the calculation of the nonlinear observation impact (solid circles and stars). The adjoint estimation of observation impacts are approximately 20-30% larger in magnitude for the self-analysis case, with smaller differences between the two verification methods for longer forecast periods. The larger impacts with self-analysis verification are a direct result of the larger forecast error difference between the pairs of forecasts as demonstrated in Figure 10. The error difference between the two forecasts includes both the true error growth (ie the difference between the dashed lines in Figure 10) and also the illusory error growth that is actually the decrease in correlation of the self-analysis verification with longer forecasts. The self-analysis estimate of forecast error is most incorrect at the analysis time, with substantially inflated error growth rates during the initial forecast period.

The net adjoint impact in the OSSE case can be compared in Figure 1 for self-analysis (grey bars) and NR (white bars) verification for the 24 hour forecast. For radiance types, the self-analysis verification impacts are of similar or greater magnitude for all instruments except for MHS. For conventional types, the self-analysis verificaiton impacts are similar or greater for all types except for rawinsonde humidities. Wind observations in particular tend to have considerably greater impact for self-analysis verification than for NR verification.

The normalized adjoint estimated observation impacts calculated using self-analysis verification are shown in Figure 11, where the normalization is against the 24 hour impact for each data type. Figure 11 may be compared with the impacts calculated using the NR verification in Figure 6. For most data types, the progression of observation impact with forecast length is similar for both choices of verification. There are however a few data types with quite different magnitudes or behavior, in particular rawinsonde winds and aircraft winds in the NHEX region, AIRS, HIRS4, and GPSRO in the extratropics and CrIS in the SHEX region. With NR verification, these observations have small beneficial impacts for short forecast lengths and increasing magnitude observation impacts for longer forecasts. However for self-analysis verfication, the short term forecast impacts are overestimated, with decreasing or steady magnitude of impact for longer forecasts. Rawinsonde temperatures and CrIS in the NHEX region show a less pronounced version of this behavior, with some inflation of observation impact magnitude for short forecasts with self-analysis verification.

This discrepancy in the magnitude of the adjoint estimation of observation impact only for certain data types and regions raises several questions. Aircraft and rawinsonde winds both demonstrate inflation of short term forecast impacts with self-analysis verification, however AMVs are not as prone to the overestimation of observation impacts. Rawinsonde and to a certain extent, aircraft are heavily weighted by the DAS and have relatively large per-observation contribution to the analysis increment, especially as there are few wind observations compared to temperature and radiance data. These two data types may also be expected to have impacts that are retained for more analysis cycles than many other types, as there are many fewer rawinsondes at 0600 UTC than at 0000 UTC, and aircraft observations also have a strong diurnal cycle in local observation count. Therefore the analysis state during the 0600 UTC cycle will have fewer corrections from new rawinsonde and aircraft data in the regions that were populated by observations at 0000 UTC, and the information from the 0000 UTC observations may persist longer, resulting in a more correlated estimate of forecast with the analysis for these particular observation types at short forecast times. Data types that have more frequent observations will have new information added to the next analysis cycle at 0600 UTC, and the self-analysis verification will be less correlated for short forecasts.

Unlike conventional rawinsonde and aircraft observations, radiance observations do not have a large diurnal cycle in availability. However, the HIRS4 and CrIS data types have small net impact (Figure 1) which is fairly noisy, as evidenced by the wide whiskers in Figures 6 and 11, particularly for longer forecasts. GPS-RO also lacks a diurnal cycle; however there is a known bias between the operator used to generate the synthetic GPSRO observations (ROPP) and the operator used to ingest the observations into the DAS, with a substantial bias in bending angle occurring in the upper troposphere. Necker *et al.* (2018) found that biased observations can have large impact on estimations of FSO, which may contribute to the overinflation of GPSRO impacts for short forecasts.

The fraction of observations with beneficial impact calculated using self-analysis verification is shown in Figure 12. Compared to the NR verification in Figure 7, the short term forecast percentages are higher for all data types, with the 6-hour forecast percentages for conventional data types being particularly large, as high as 70% for rawinsonde winds in the Tropics. Jung *et al.* (2013) found percentages of beneficial observations of 60-70% for 6-hour forecast impacts using self-analysis verification, although their fractions of beneficial impacts for the 24-hour forecast timeframe only decreased to 60-66%, while the fractions found here at 24 hours are in the range of 50-55%. Kotsuki *et al.* (2019) found fraction of beneficial observations near 59% with self-analysis at 6 hr, and 56% at 12 hours.

As in Section 3, the impacts of select subsets of observations can be traced to different forecast times. This is of particular interest as it pertains to the Proactive Quality Control (PQC; Chen and Kalnay 2019) method in which the 10% most detrimental observations as determined by a 6-hr ensemble forecast are omitted in an attempt to improve the analysis quality and forecast skill. Self-analysis is used with PQC to determine which observations have the worst impacts. While the adjoint operates differently from the PQC methods, the self-analysis incestuousness issue can still be evaluated here.

Figure 13 compares the behaviors of several different subsets of observations from the 0000 UTC cycle time for four data types for the month of July. The subset of detrimental observations having 6 hr forecast impacts that are 0.5σ greater than the mean is approximately 10% of the total dataset. The progression of per-observation impacts for the 10% most detrimental 6 hr forecast observations as calculated using the NR fields for verification is shown with the heavy solid line, and this subset of observations will be referred to as DETNR. A similar progression of per-observation impact is shown for the 10% of most detrimental observations as determined using the self-analysis as verification is shown by the thin dashed line, and this subset of observations will be referred to at DETANA. Approximately 35-45% of the same observations are in both DETNR and DETANA.

The estimated impacts of DETANA using the NR for forecast error verification (thin solid line) and self-analysis for verification (dashed line) are fairly close for AMSU-A and AIRS, with largest discrepancy for MHS. At short forecast times, the DETANA observation subset is clearly less detrimental than the DETNR subset, but at longer forecast times, these subsets have net impacts that become more similar in magnitude, even though many of the observations in the DETANA subset are incorrectly assigned. The dash-dot lines in Figure 13 show the NR-verified per-observation impacts of the observations that are in both DETNR and DETANA (heavy dash-dot) and the observations that are in DETANA but not DETNR (thin dash-dot). The observations in DETANA that are also in DETNR have net impact that is strongly detrimental at the short forecast time and becomes more detrimental with longer forecast times. This implies that the self-analysis verification has some skill at identifying the detrimental observations with the greatest magnitude impacts. However, the observations that are in DETANA but not DETNR actually have net per-observation impact that is beneficial at short forecast times, becoming weakly detrimental at longer forecast times. This illustrates the difficulty in identifying observation impacts at short forecast times when relying on self-analysis verification.

5. Conclusions

Observation impacts on forecast skill are dependent upon the forecast error evolution during the forward model integration. FSO allows for studying the impact on forecast error resulting from small changes in initial conditions due to the the ingestion of observations, regardless of model errors. Uncertainties in the observations also impact the data assimilation cycle and thus the verifications typically used to evaluate forecast errors. When self-analysis verification is used, the incestuousness of the verification method distorts both the estimates of forecast error and the forecast error growth rate in a way that is nonlinear with forecast length. At the 6-hour forecast, the self-analysis verification grossly underestimates the total forecast error, but overestimates the forecast error growth, particularly during the first 6-12 hours of the forecast period. As the forecast lengthens to 48 hours, the distortion of the forecast error estimate by self-analysis verification is minimal, and the forecast error growth rate is only slightly overestimated.

It is not clear that an optimal forecast length for calculation of FSO exists for an operational setting where only self-analysis verification is available. At the 12-24 hour forecast length range, the FSO estimate of observation impact with self-analysis verification (open stars in Figure 4) is actually quite close to the true nonlinear observation impact verified with the NR (solid black circles), even more so than the FSO estimate using NR verification. However, this apparent veracity is more of a "lucky guess" achieved for the wrong reasons and not because the FSO with self-analysis verification is more accurate.

There are some regional variations in the progression of observation impact with forecast time that reflect the different types of model error and physical and dynamical processes that lead to forecast error growth. In the extratropics, many observation types show observation impacts that increase in magnitude with longer forecast lengths. This might be expected with errors related to baroclinic processes that have intrinsic timescales of several days. In contrast, in the Tropics, there are many observation impacts that do not substantially increase with forecast length, and may even decrease. These errors may have short timescales of growth, such as due to convective or other physical processes, and model errors may grow rapidly and erase the useful information added by observations.

Moisture-based data such as in-situ humidity observations and the microwave humidity sounder (MHS) show similar behaviors globally. These data have initially large magnitude observation impacts and high percentages of beneficial observations, both of which decrease with longer forecast times. This combination of traits is strongly suggestive of large background errors in the humidity field due to fast acting model errors. Large background errors present the opportunity for the observations to perform a substantial amount of "work" in correcting the analysis field. The rapid decrease in impacts with forecast time indicates that these initial improvements are not maintained into the forecast beyond the first day of integration, presumably because of types of error growth that cannot be corrected by the observations (i.e., model error).

One of the major omissions from the OSSE framework is the lack of realistic model error and observation biases. Necker *et al.* (2018) and Kotsuki *et al.* (2019) have found that biases can have large effects when calculating FSO. The GMAO OSSE is not completely devoid of biases - there are some model biases that result from differences in model physics between the G5NR and the forecast model. There are also some observation biases that are introduced through the observation operators, such as known biases between the ROPP operator used for simulating GPSRO bending angles and the GSI operator used to ingest the observations. The bias correction is also allowed to act, even though the observations do not have explicitly added biases. Thus, the bias correction may attempt to “correct” what it sees as observation errors but what are in fact model biases. It is likely that some of the difference between the magnitude of the Real versus OSSE FSO calculations is due to the lack of biases in the OSSE.

When considering the NR as verification, biases in the observations and biases in the model error will both tend to decrease the beneficial impact of observations. Observation bias will tend to introduce analysis errors, unless the biases are removed by bias correction. Model biases will tend to remove useful information from assimilated observations and shorten the timescale on which observations provide positive impacts.

The situation is more complex with self-analysis verification, as biases that result in analysis bias can affect the calculation of FSO. When an observation bias is ingested by the DAS but is corrected by other data types, the analysis field may be minimally impacted, and the bias will cause a decrease in the beneficial impact of that observation, as with NR verification. Alternatively, when observation biases reinforce existing analysis biases, observations may be seen as having more beneficial impact due to the bias. If the model has a bias that is not corrected by observations, so that the analysis field is similarly biased, then the unbiased observations may be seen as having a less beneficial impact when self-analysis verification is used. When bias correction is implemented where the model is assumed to be unbiased and all biases are assigned to observations, a model bias will be present in the analysis field and the observations themselves will be adjusted to include a similar bias, and the beneficial impact of these adjusted observations might be overinflated with self-analysis verification.

The impact of any individual observation will follow a progression as the forecast integrates forward in time that depends upon the growth and decay of the background state errors that are adjusted by ingestion of the observation by the DAS. In a sampling of observations tested here, less than a third of the observations that have the strongest beneficial impacts on the 6 hour forecast maintained that strong impact to the 48 hour forecast time. This progression of observation impacts is further complicated in an operational setting where only self-analysis verification is available. The identification of particular observations with strongly beneficial or detrimental impacts is particularly challenging for short forecast lengths, where the incestuousness of self-analysis verification interferes with the accurate estimation of observation impacts.

There are two concerns for methods such as PQC which rely on identifying detrimental observations in 6-hour forecasts. First, there is the question of whether observations which are detrimental at 6 hours are representative of the observations that are detrimental at longer forecasts. Our results show that while the net impact of the most detrimental observations at 6 hours remains detrimental up to 48 hours, many of these individual observations have beneficial impacts particularly at and beyond 24 hours. Also, only a fraction of the observations with the most detrimental impact at 48 hours have detrimental impact at 6 hours. Second, there is a concern for accurately identifying the most detrimental observations at 6 hours given the lack of available independent verification data. When using self-analysis verification, the success rate at accurately selecting the most detrimental observations at 6 hrs is approximately 40%. For a method such as PQC to have a chance to work, it is fundamental for errors to be defined with respect to verification fields independent from the data assimilation cycle.

Acknowledgement

Support for this project was encouraged by Steven Pawson and provided by NASA/GMAO core funding. Resources supporting this work were provided by the NASA High-End Computing (HEC) Program through the NASA Center for Climate Simulation (NCCS) at Goddard Space Flight Center. The software for simulating GPSRO observations was provided by the Radio Occultation Processing Package (ROPP) of the Radio Occultation Meteorology (ROM) Satellite Applications Facility (SAF) of EUMETSAT, with the assistance of Sean Healy at ECMWF. The authors wish to thank Olaf Stiller and an anonymous reviewer for their helpful comments.

References

Baker N, Daley R. 2000. Observation and background adjoint sensitivity in the adaptive observation targeting problem. *Quart. J. Roy. Meteor. Soc.* **126**: 1431–1454.

Barahona D, Molod A, Bacmeister J, Nenes A, Gettelman A, Morrison H, Phillips V, Eichmann A. 2014. Development of two-moment cloud microphysics for liquid and ice within the NASA Goddard Earth Observing System Model (GEOS-5). *Geoscience Model Development* **7**: 1733–1766. Doi: 10.5194/gmd-7-1733-2014.

Cardinali C. 2018. Forecast sensitivity observation impact with an observation-only based objective function. *Quart. J. Roy. Meteor. Soc.* **144**: 2089–2098. DOS: 10.1002/qj.3305.

Chen TC, Kalnay E. 2019. Proactive quality control: observing system simulation experiments with the Lorenz '96 model. *Mon. Wea. Rev.* **147**: 53–67. Doi.org/10.1175/MWR-D-18-0138.1.

Culverwell I, Lewis H, Offiler D, Marquardt C, Burrows C. 2015. The radio occultation processing package, ROPP. *Atmos. Meas. Tech.* **8**: 1887–1899. Https://doi.org/10.5194/amt-8-1887-2015.

Daescu D. 2009. On the deterministic observation impact guidance: a geometrical perspective. *Mon. Wea. Rev.* **137**: 3567–3574.

Ehrendorfer M. 2007. A review of issues in ensemble-based Kalman filtering. *Meteorologische Zeitschrift* **16**: 795–818.

Ehrendorfer M, Errico R. 1995. Mesoscale predictability and the spectrum of optimal perturbations. *J. Atmos. Sci.* **52**: 3475–3500.

Errico R, Ehrendorfer M, Raeder K. 2001. The spectra of singular values in a regional model. *Tellus* **53A**: 317–332.

Errico R, Privé N, Carvalho D, Sienkiewicz M, Akkraoui AE, Guo J, Todling R, McCarty W, Putman W, da Silva A, Gelaro R, Moradi I. 2017. Description of the GMAO OSSE for Weather Analysis software package: Version 3. Technical Report 48, National Aeronautics and Space Administration. NASA/TM-2017-104606.

Errico RM, Carvalho D, Privé NC, Sienkiewicz M. 2020. Simulation of atmospheric motion vectors for an observing system simulation experiment. *J. Atmos. Ocean Tech.* **37**: 489–505. Doi:10.1175/JTECH-D-19-0079.1.

Errico RM, Privé NC. 2018. Some general and fundamental requirements for designing observing system simulation experiments (osses). *WMO Rep. WWRP 2018-8* : 33 pp.URL https://www.wmo.int/pages/prog/arep/wwrp/new/documents/Final_WWRP_2018_8.pdf.

Errico RM, Yang R, Privé N, Tai KS, Todling R, Sienkiewicz M, Guo J. 2013. Validation of version one of the Observing System Simulation Experiments at the Global Modeling and Assimilation Office. *Quart. J. Roy. Meteor. Soc.* **139**: 1162–1178. Doi: 10.1002/qj.2027.

Gelaro R, Langland R, Pellerin S, Todling R. 2010. The THORPEX observation impact intercomparison experiment. *Mon. Wea. Rev.* **138**: 4009–4025.

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

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

Han Y, van Delst P, Liu Q, Weng F, Yan B, Treadon R, Derber J. 2006. JCSDA Community Radiative Transfer Model (CRTM) - version 1. OAA Tech. Report 122.

Holdaway D, Errico R, Gelaro R, Kim J. 2014. Inclusion of linearized moist physics in NASA's Goddard Earth Observing System data assimilation tools. *Mon. Wea. Rev.* **142**: 414–433.

Hotta D, Chen T, Kalnay E, Ota Y, Miyoshi T. 2017. Proactive QC: a fully flow-dependent quality control scheme based on EFSO .

Jung B, Kim H, Auligné T, Zhang X, Huang X. 2013. Adjoint-derived observation impact using WRF in the western North Pacific. *Mon. Wea. Rev.* **141**: 4080–4097. Doi.org/10.1175/MWR-D-12-00197.1.

Kleist D, Parrish D, Derber J, Treadon R, Wu WS, Lord S. 2009. Introduction of the GSI into the NCEP global data assimilation system. *Weather and Forecasting* **24**: 1691–1705.

Kotsuki S, Kurosawa K, Miyoshi T. 2019. On the properties of ensemble forecast sensitivity to observations. *Quart. J. Roy. Meteor. Soc.* **145**: 1897–1914. Doi.org/10.1002/qj.3534.

Langland R, Baker N. 2004. Estimation of observation impact using the NRL atmospheric variational data assimilation adjoint system. *Tellus-A* **56**: 189–201. Doi.org/10.1111/j.1600-0870.2004.00056.x.

Lawless A. 2010. A note on the analysis error associated with 3D-FGAT. *Quart. J. Roy. Meteor. Soc.* **136**: 1094–1098. Doi.org/10.1002/qj.619.

Legras B, Vautard R. 1996. A guide to Liapunov vectors. In: *Predictability, Volume I.* European Centre for Medium-Range Weather Forecasts,, pp. 143–156.

Lorenc A, Marriott R. 2014. Forecast sensitivity to observations in the Met Office Global numerical weather prediction system. *Quart. J. Roy. Meteor. Soc.* **140**: 209–224. Doi: https://doi.org/10.1002/qj.2122.

Massart S, Pajot B, Piacentini A. 2010. On the merits of using a 3D-FGAT assimilation scheme with an outer loop for atmospheric situations governed by transport. *Mon. Wea. Rev.* **138**: 4509–4522. Doit: 10.1175/2010MWR3237.1.

Necker T, Weissman M, Sommer M. 2018. The importance of appropriate verification metrics for the assessment of observation impact in a convection-permitting modelling system. *Quart. J. Roy. Meteor. Soc.* **144**: 1667–1680.

Privé N, Errico R, Tai KS. 2013a. The influence of observation errors on analysis error and forecast skill investigated with an observing system simulation experiment. *J. Geophys. Res.* **118**: 5332–5346. Doi: 10.1002/jgrd.50452.

Privé N, Errico R, Tai KS. 2013b. Validation of forecast skill of the Global Modeling and Assimilation Office observing system simulation experiment. *Quart. J. Roy. Meteor. Soc.* **139**: 1354–1363. Doi: 10.1002/qj.2029.

Rienecker M, Suarez M, Todling R, Bacmeister J, Takacs L, Liu HC, Gu W, Sienkiewicz M, Koster R, Gelaro R, Stajner I, Nielsen J. 2008. The GEOS-5 data assimilation system - documentation of versions 5.0.1, 5.1.0 and 5.2.0. Technical Report 27, NASA.

Todling R. 2013. Comparing two approaches for assessing observation impact. *Mon. Wea. Rev.* **141**: 1484–1505.

Trémolet Y. 2008. Computation of observation sensitivity and observation impact in incremental variational data assimilation. *Tellus* **60**: 964–978.

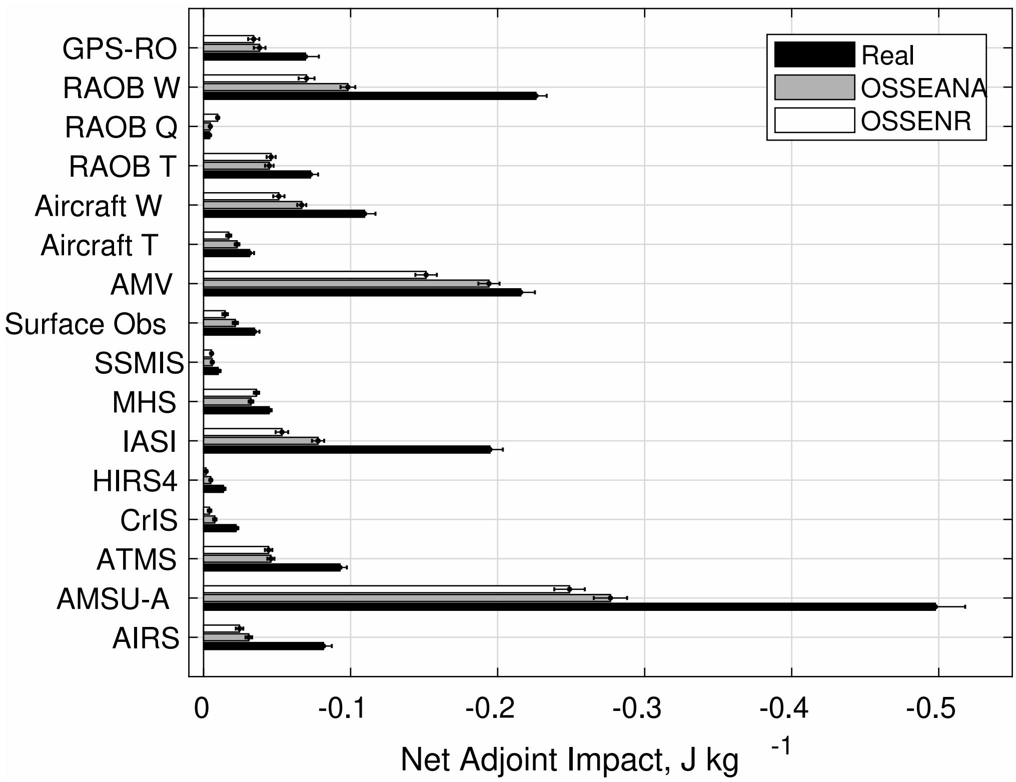

Figure 1. Global net FSO estimated observation impact on total wet error energy (Equation (1)) at the 24 hour forecast for select data types ($J\ kg^{-1}$), mean over two month period. Black, Real case with self-analysis verification; grey, OSSE case with self-analysis verification; white, OSSE case with NR verification. Negative values indicate a reduction in the 24-hour forecast error, note scale and reverse direction of abscissa. Whiskers indicate 95% confidence intervals.

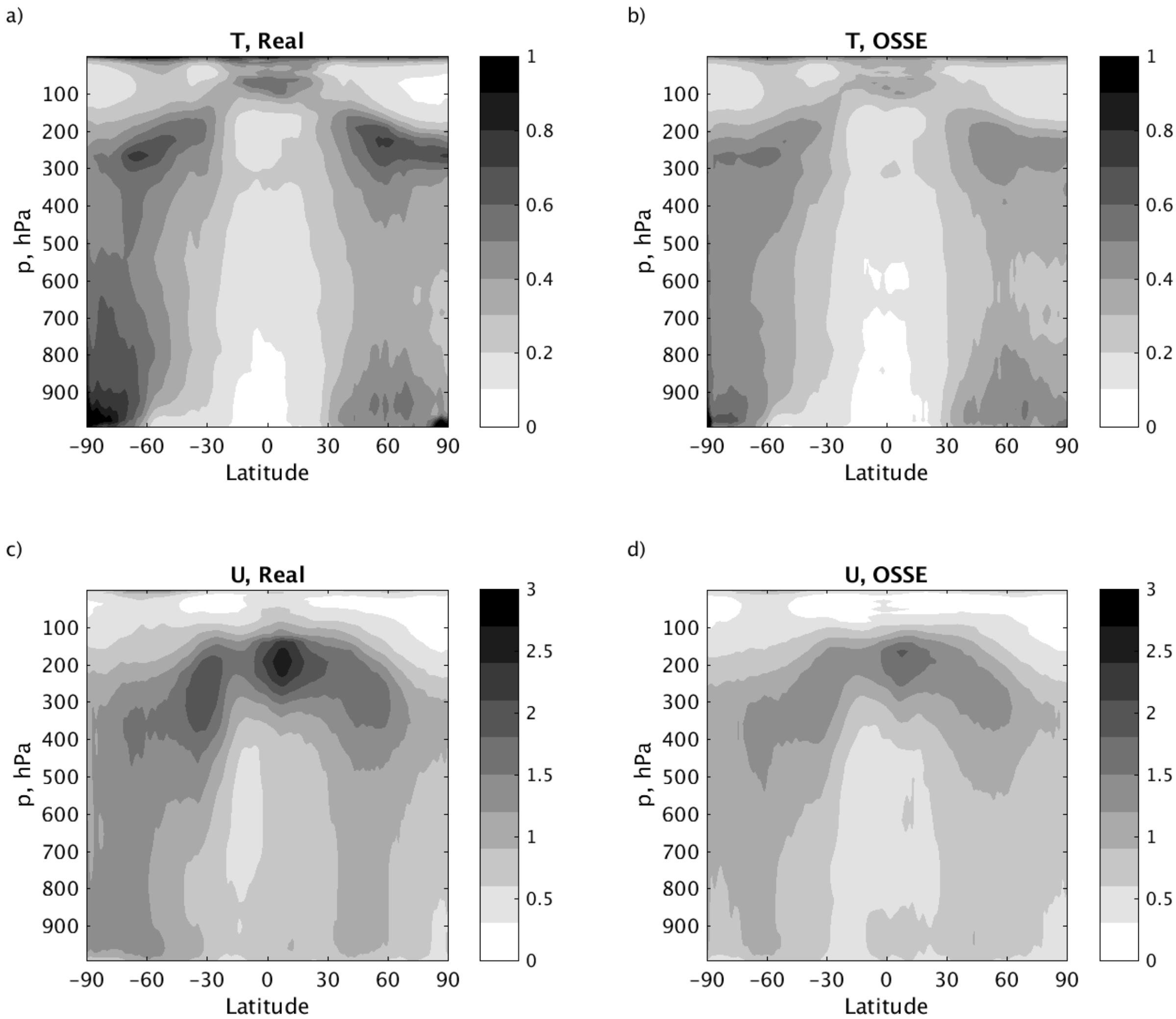

Figure 2. Zonal mean temporal root mean square of analysis minus background fields (A-B) for July and August. a, b) temperature (K); c, d) zonal wind (m s^{-1}); a,c) real data (2015); b, d) OSSE (2006).

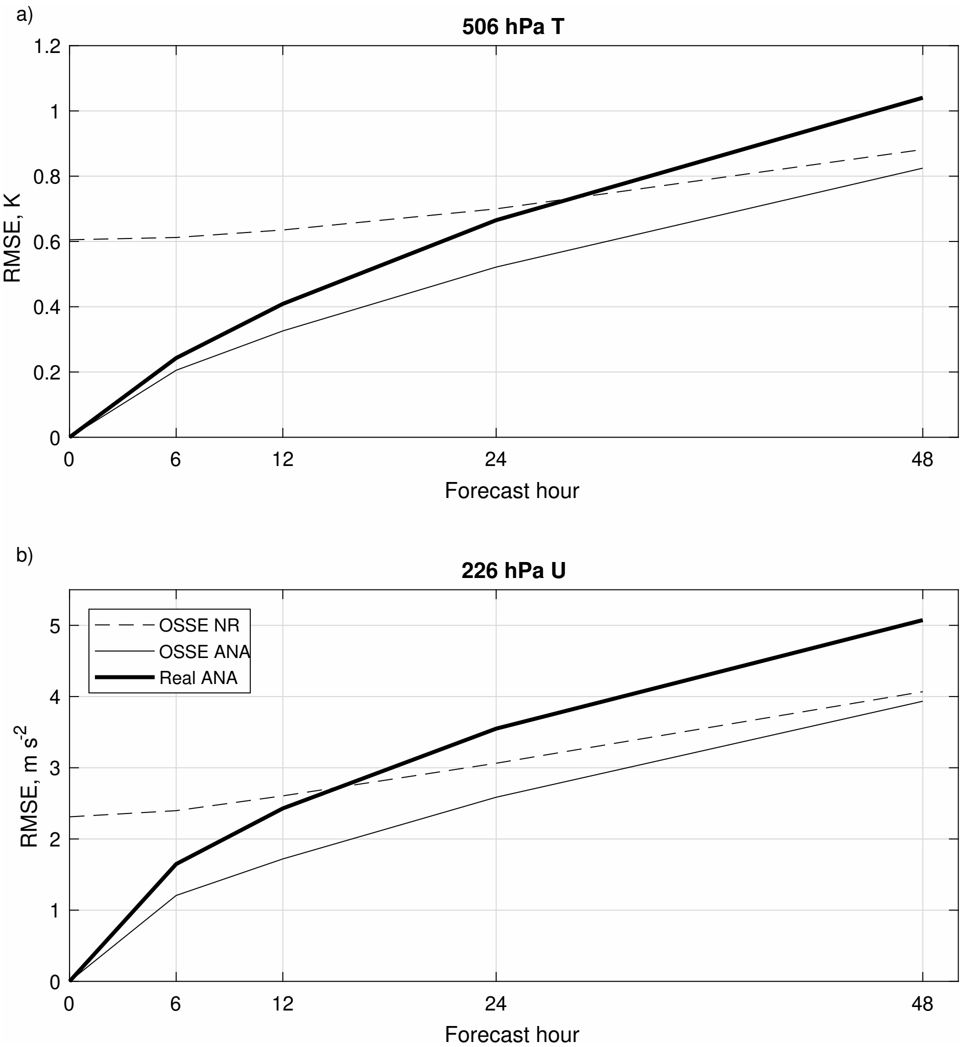

Figure 3. Areal mean of the root-temporal mean-square forecast error for July and August as a function of forecast length. Heavy solid line, real data with self-analysis verification; thin solid line, OSSE with self-analysis verification; thin dashed line, OSSE with NR verification. a) temperature on the 506 hPa model surface (K); b) zonal wind on the 226 hPa surface (m s^{-1}).

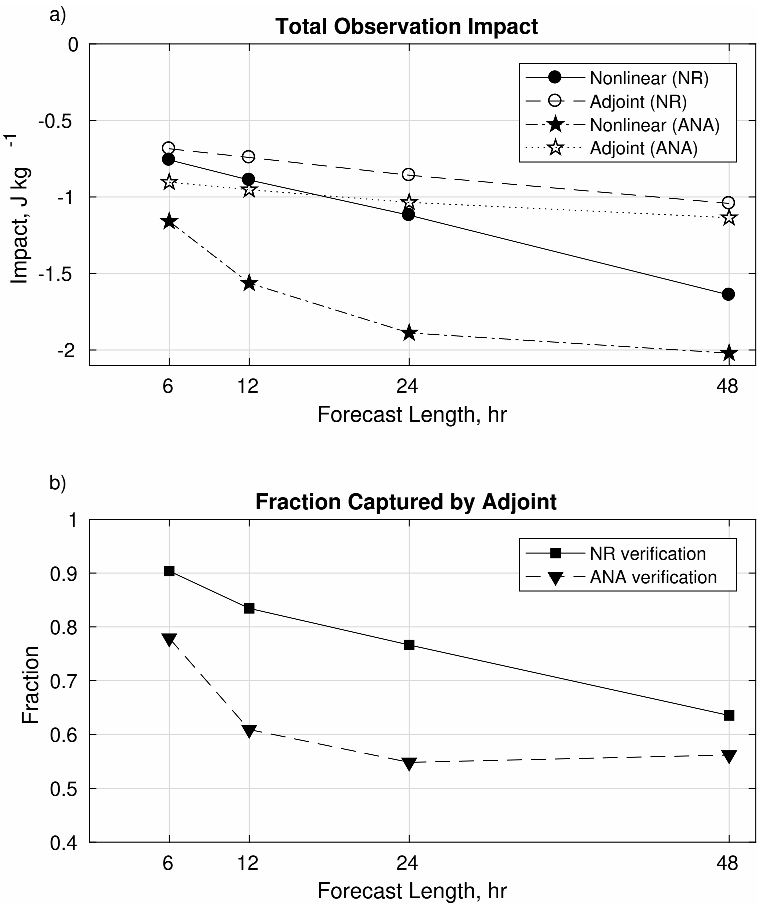

Figure 4. Total observation impact calculated as a function of forecast length for the nonlinear difference between forecast pairs (filled shapes) and the adjoint estimate of the total impact (open shapes). Circles, NR verification; stars, self-analysis verification. b) Fraction of the nonlinear observation impact captured by the adjoint as a function of forecast length.

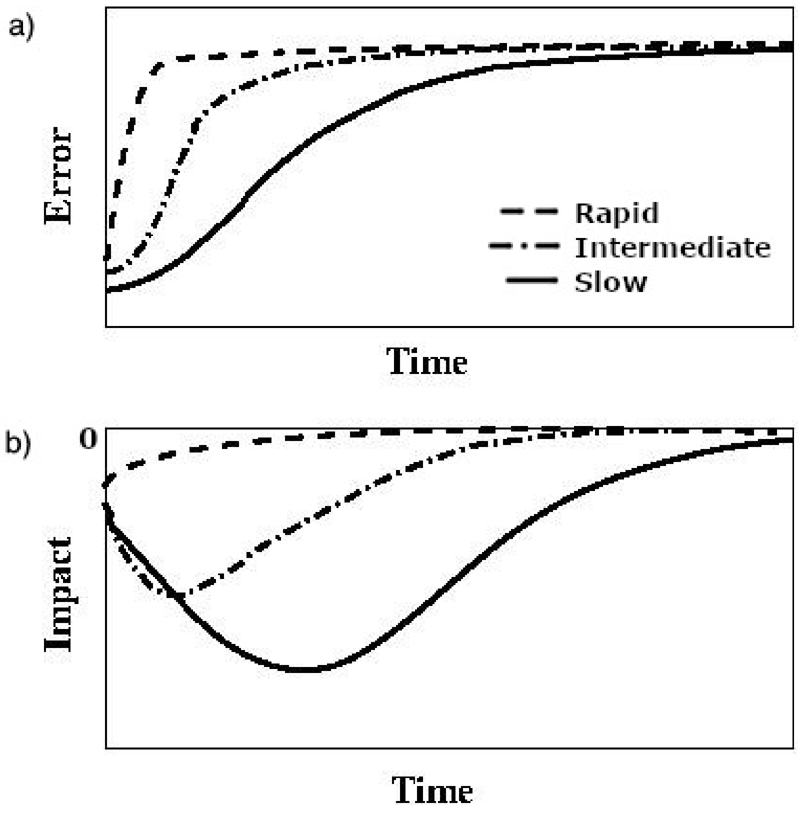

Figure 5. Schematic illustration of the evolution of forecast errors and observation impacts with forecast length. The lines represent cases with different rates of growth and saturation of error. The dashed line indicates the most rapid error growth and saturation, the solid line represents more gradual error growth; and the dash-dot line represents an intermediate rate of error growth. a) The growth of the error norm associated with errors having different timescales of saturation; b) the observation impacts that project onto these corresponding errors, drawn with negative impacts for consistency with other figures.

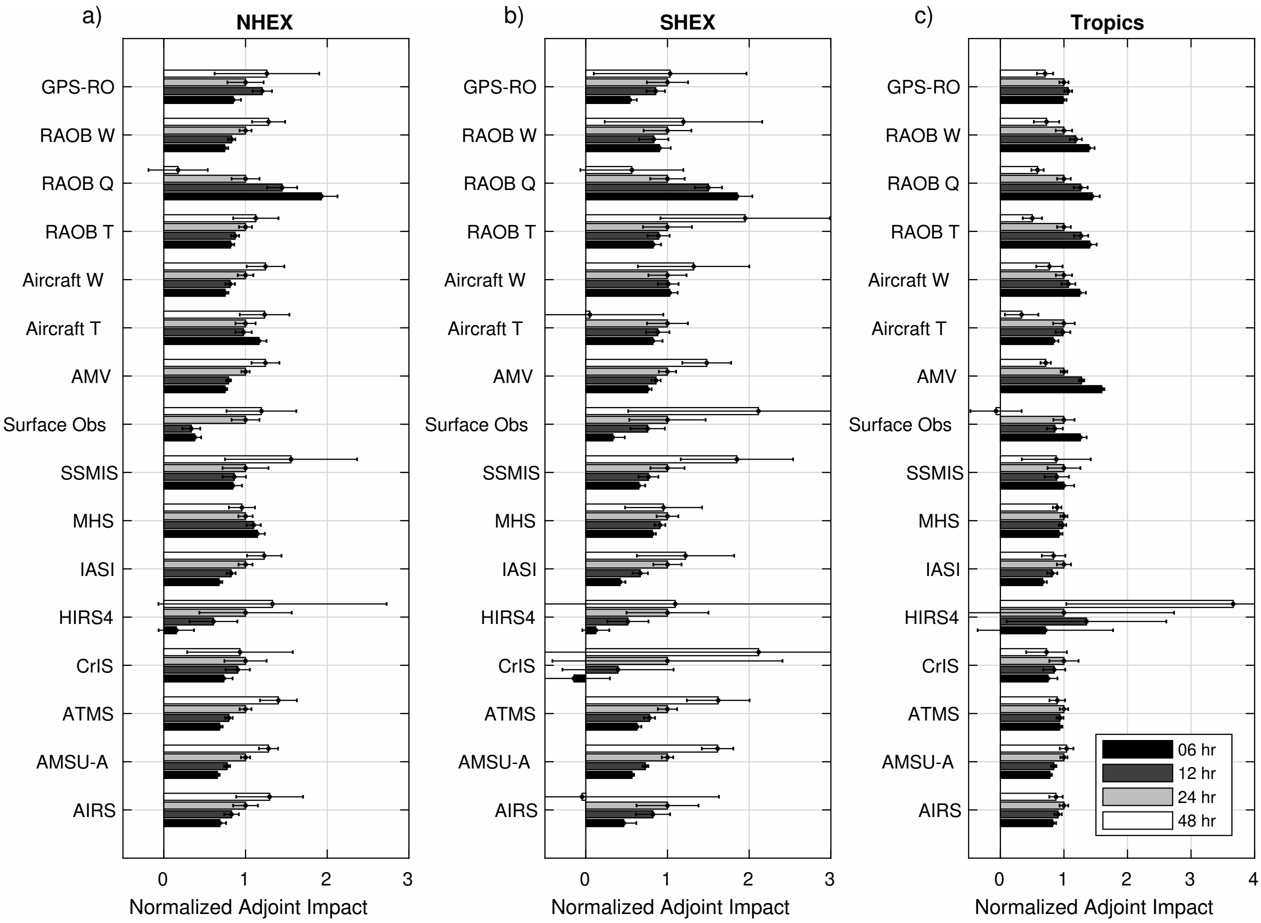

Figure 6. Normalized adjoint estimated observation impact on total wet energy norm per cycle for select data types relative to 24-hour observation impacts, mean over two month period, for forecasts of length 6, 12, 24, and 48 hours. NR verification. a) NHEX region; b) SHEX region; c) Tropics region. Whiskers indicate 95th percentile confidence interval.

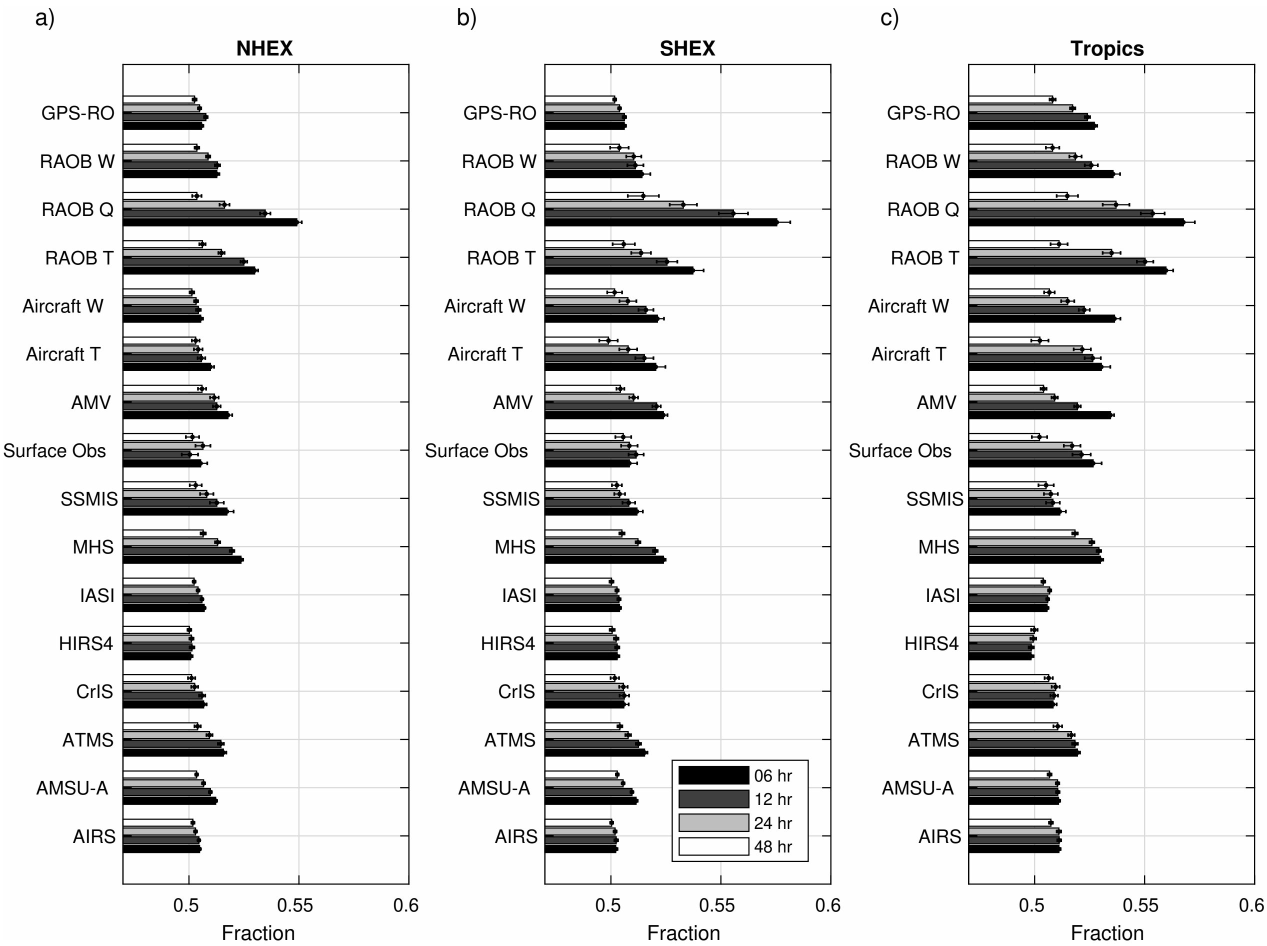

Figure 7. Fraction of observations with negative beneficial impact on total wet energy for select data types, mean over two month period, for forecasts of length 6, 12, 24, and 48 hours, using NR verification. a) NHEX region; b) SHEX region; c) Tropics region. Whiskers indicate errorbars for 95th percentile confidence interval.

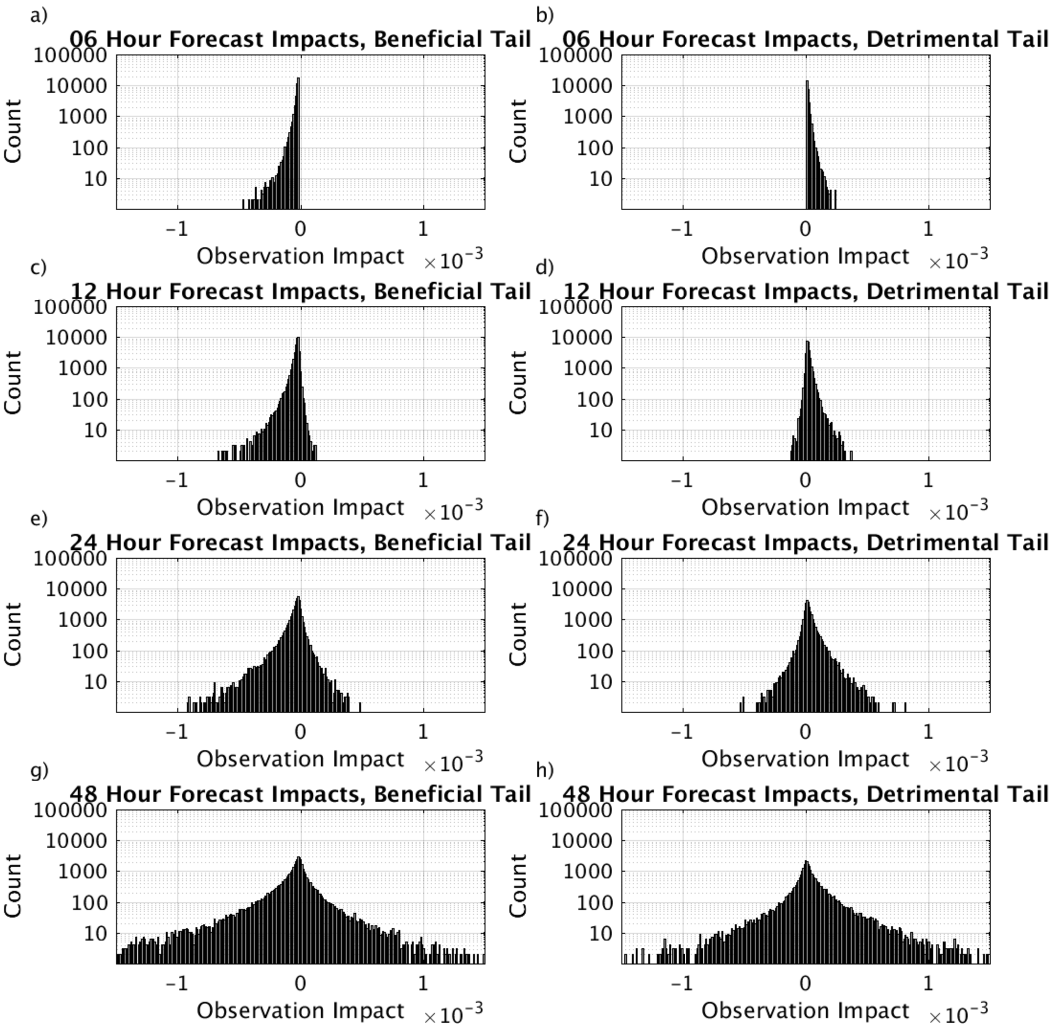

Figure 8. Histogram of counts of observations (ordinate, log scale) according to their observation impacts (abscissa, NR verification) for AMSU-A NOAA-19 observations, cumulative for 0000 UTC observations from 2 July to 30 July, NR verification. Negative (positive) tail of the distribution at 06 hours selected for observation impacts less (greater) than 2.5 standard deviations from the mean. Left, negative (beneficial) tail at 06 hr forecast; right, positive (detrimental) tail at 06 hr forecast. a, b) 06 hr forecast; c, d) 12 hour forecast; e, f) 24 hour forecast; g, h) 48 hour forecast.

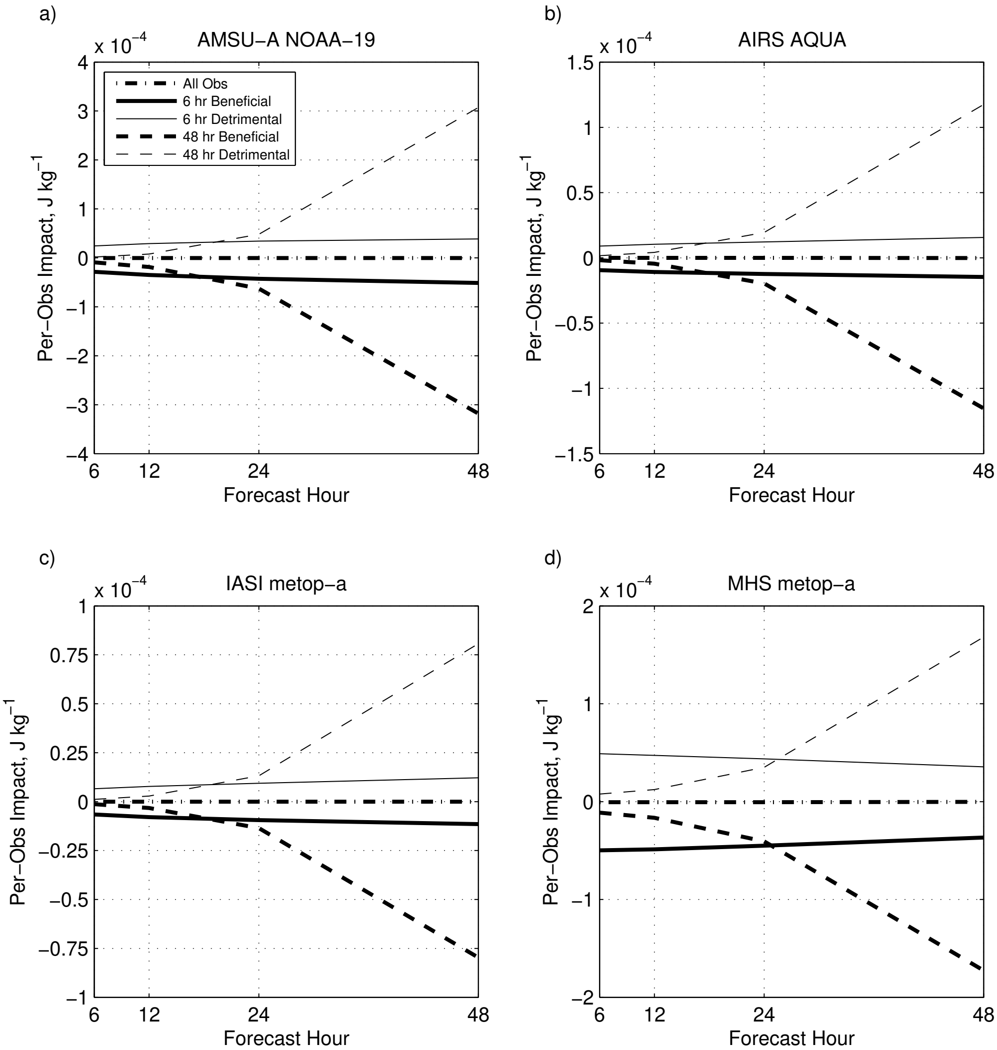

Figure 9. Per-observation impacts for subsets of observations as a function of forecast time, NR verification, cumulative dataset for the month of 0000 UTC forecasts in July. a) AMSU-A NOAA-19; b) AIRS AQUA; c) IASI metop-a; MHS metop-a. Heavy lines: negative (positive, thin lines) tail of the distribution at 06 hours selected for observation impacts less (greater) than 2.5 standard deviations from the mean; similar calculations are made for the negative (positive) tail of the 48 hour forecast observation impacts, dashed lines. Set of all observations, heavy dash dot line.

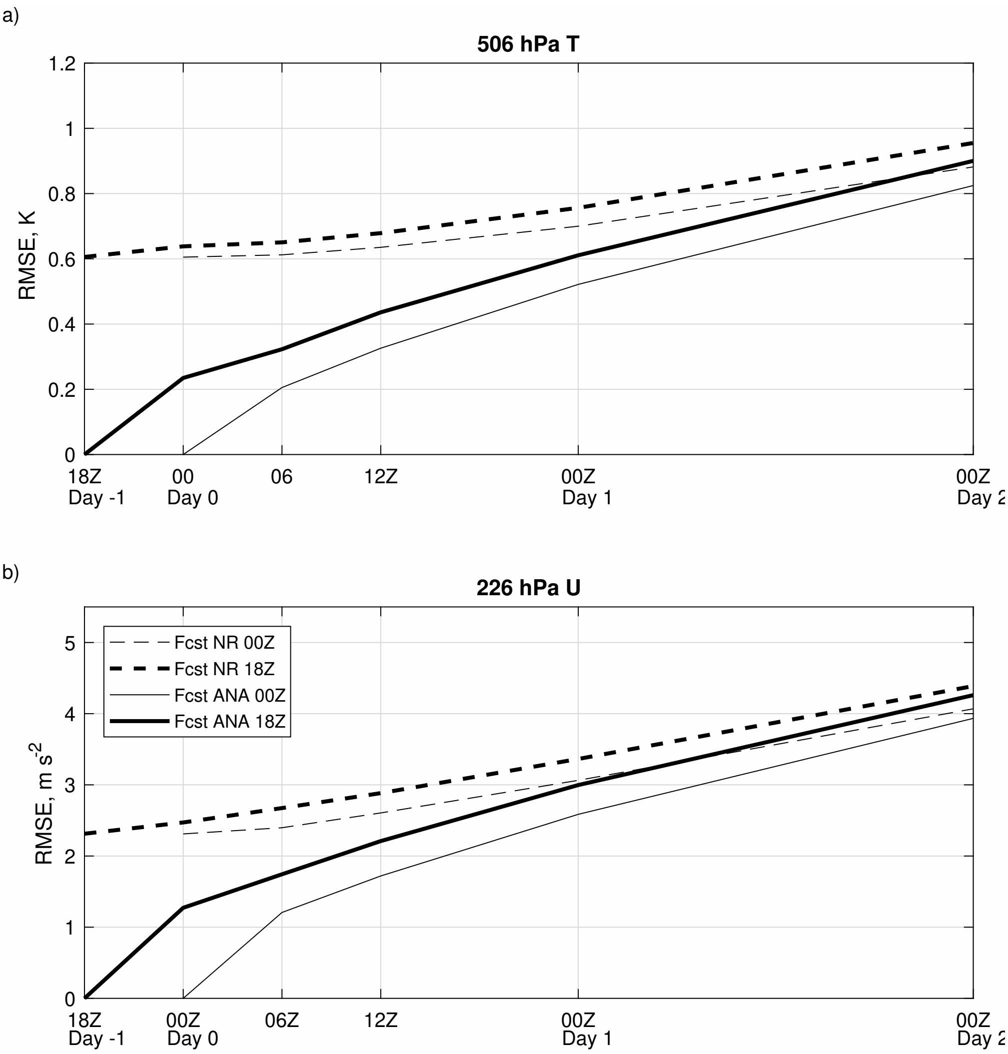

Figure 10. Areal mean of the RMSE forecast error for July and August as a function of forecast length. Solid lines, self-analysis verified forecast starting from 0000 UTC (thin) and 1800 UTC (thick). Dashed lines, NR verified forecast starting from 0000 UTC (thin) and 1800 UTC (thick). a) temperature on the 506 hPa model surface (K); b) zonal wind on the 226 hPa surface (m s^{-1}).

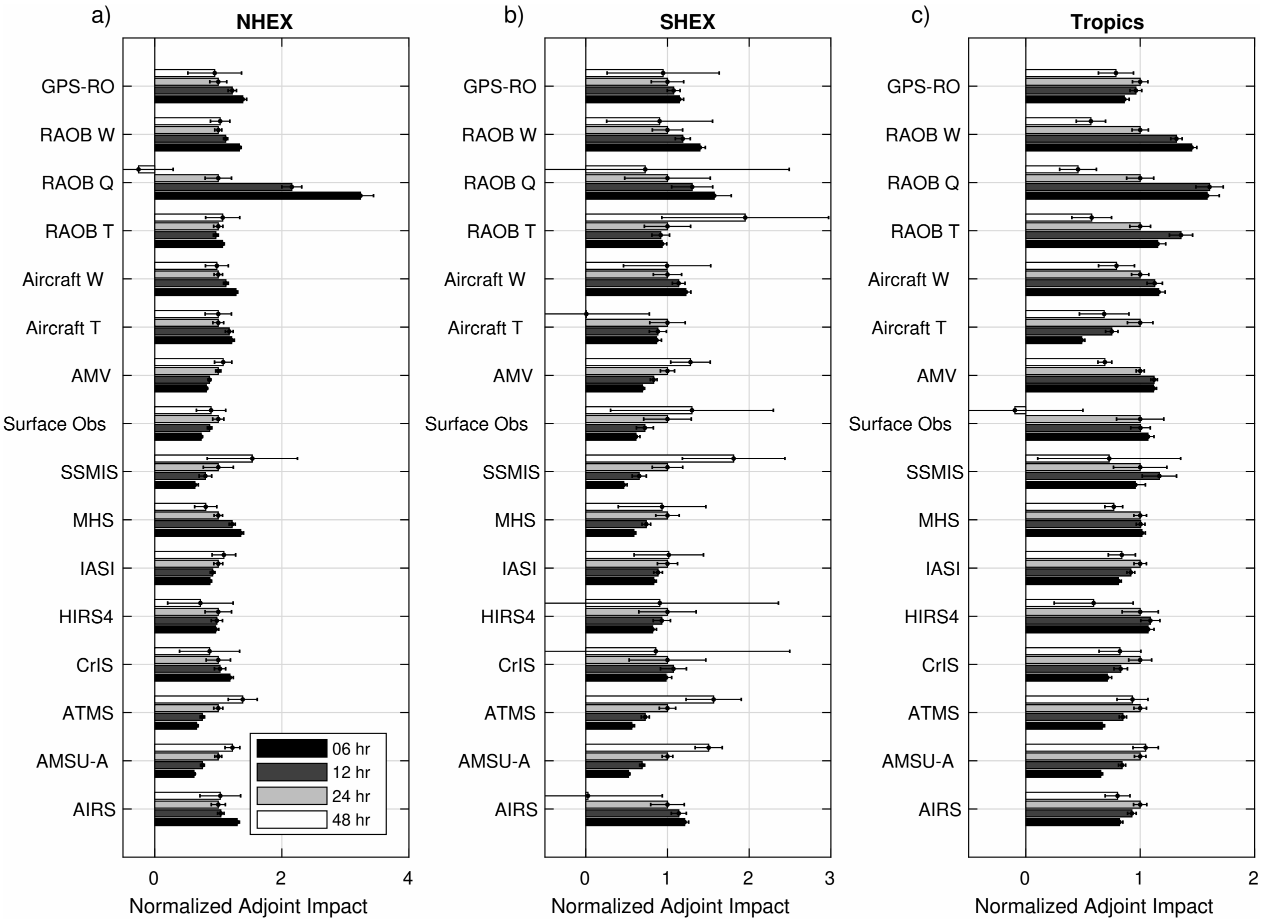

Figure 11. Normalized adjoint estimated observation impact on total wet energy per cycle for select data types ($J\ kg^{-1}$), mean over two month period, for forecasts of length 6, 12, 24, and 48 hours, using self-analysis verification, normalized by 24 hour forecast impacts. a) NHEX region; b) SHEX region; c) Tropics. Error bars indicate 95% confidence intervals.

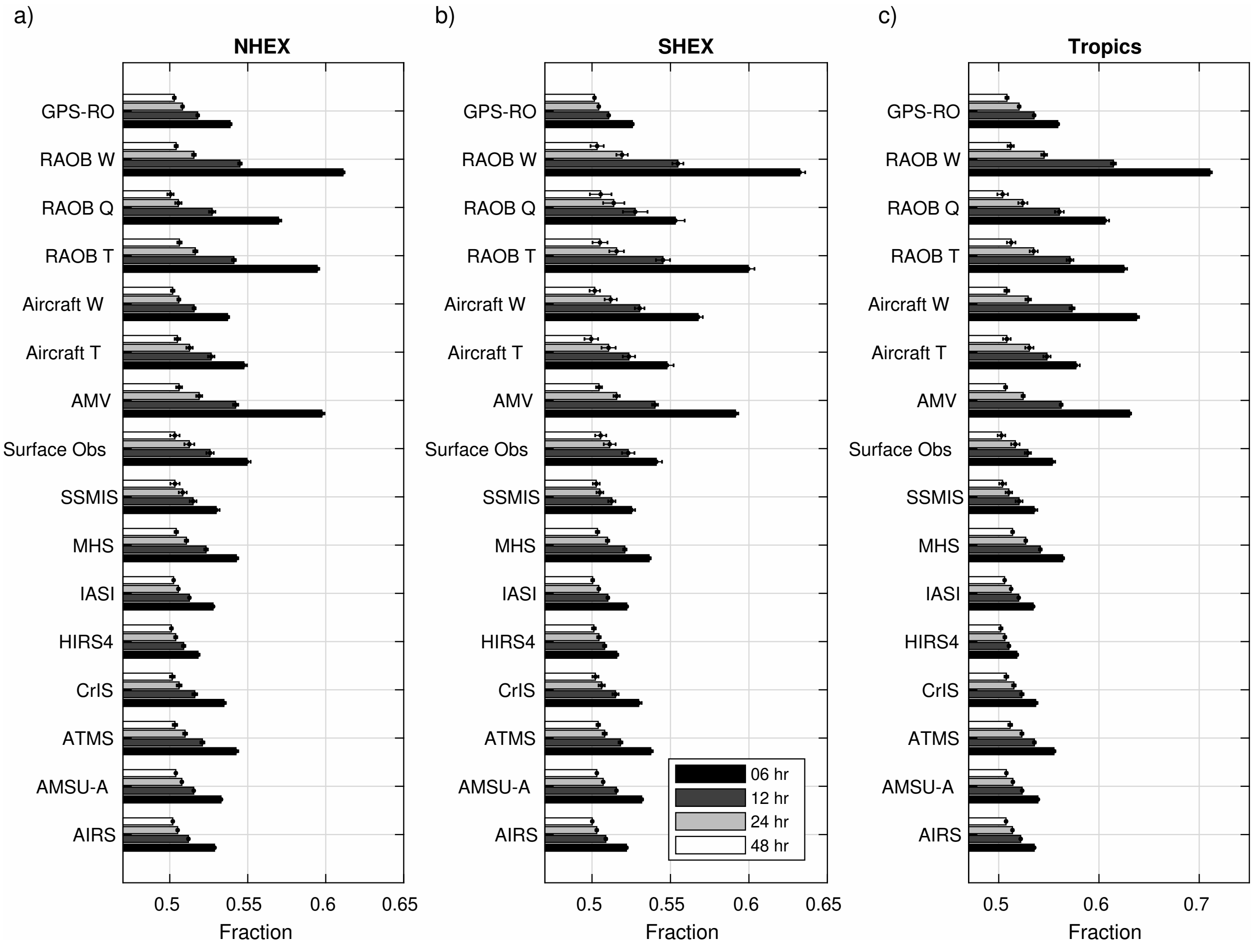

Figure 12. Fraction of observations with negative (beneficial) impact on total wet energy for select data types, mean over two month period, for forecasts of length 6, 12, 24, and 48 hours, using self-analysis verification. Note different abscissa scales between panels and in comparison to Figure 7. a) NHEX region; b) SHEX region; c) Tropics region. Error bars indicate 95% confidence intervals.

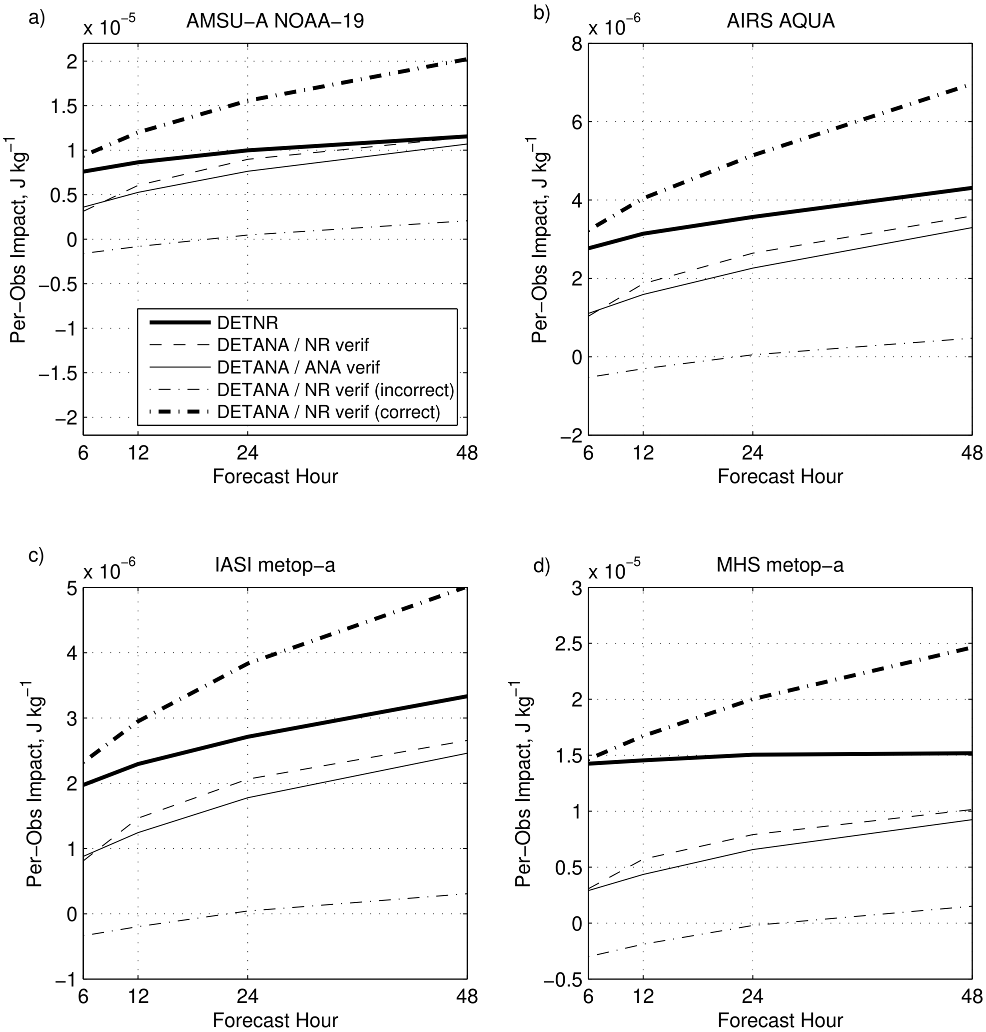

Figure 13. Per-observation impacts for subsets of observations as a function of forecast time for several data types, cumulative dataset for 0000 UTC forecasts for the month of July. Solid heavy line, 10% most detrimental observations determined with NR verification at 6 hrs (DETNR) and verified with the NR for longer forecasts; thin solid line, 10% most detrimental observations at 6 hr as verified with self-analysis (DETANA) with impacts calculated with NR verification; thin dashed line, DETANA with impacts verified by self-analysis. Thin dash-dot line, incorrectly assigned members of DETANA with NR verified impacts; heavy dash dot line, correctly assigned members of DETANA with NR verified impacts. a) AMSU-A NOAA-19; b) AIRS AQUA; c) IASI metop-a; d) MHS metop-a.