

The importance of simulated errors in observing system simulation experiments

By NIKKI C. PRIVÉ^{*1}, RONALD M. ERRICO², and WILL MCCARTY³, ¹*Goddard Earth Sciences Technology and Research, Morgan State University, Greenbelt, MD, USA;* ²*Goddard Earth Sciences Technology and Research, University Space Research Association, Greenbelt, MD, USA;* ³*National Aeronautics and Space Administration, Global Modeling and Assimilation Office, Greenbelt, MD, USA*

(Manuscript received 20 July 2020; in final form 1 February 2021)

ABSTRACT

Observing System Simulation Experiments (OSSEs) for numerical weather prediction rely on simulated observations that should include simulated observation errors in order to realistically represent the behaviour of real data. Real observations include many types of error, such as instrument error, representativeness error, and observation operator error, with some portion of this error being correlated in time and space or possibly between data types. Data assimilation systems are designed to account for random, uncorrelated errors, but are not yet adept at handling correlated errors; as a result, the correlated errors are more readily incorporated into the analysis increment by the data assimilation system than uncorrelated errors. In this work, the role of correlated observation errors in modifying the behaviour of the National Aeronautics and Space Administration Global Modeling and Assimilation Office (NASA/GMAO) OSSE framework is investigated. The effects on analysis increment, analysis error, forecast errors and observation impacts of including or neglecting correlated simulated errors is explored. The use of correlated observations for calibration and validation of the OSSE is also discussed.

Keywords: data assimilation, observing system simulation experiment, observation error

1. Introduction

Numerical weather prediction (NWP) Observing System Simulation Experiments (OSSEs) are used for evaluation of the impacts of potential new observing systems on operational forecast skill, as well as for investigation of data assimilation system (DAS) behaviour. In an OSSE, the real world is replaced with a model proxy, or Nature Run (NR), usually a long free forecast from a sophisticated general circulation model (GCM). Simulated or ‘synthetic’ observations are created using the NR fields, where the synthetic observations are intended to represent what current and future instruments would measure when deployed to observe the NR. These synthetic observations are then ingested into an operational NWP DAS with the option to generate forecasts that can be verified against the NR (Errico and Privé, 2018). The OSSE is a completely simulated framework where the ‘true’ state of the atmosphere is fully known, allowing explorations of the observing system and NWP setup that are not possible with real data. Future instrument types that do not

currently exist can be tested and various configurations compared (ex. Stoffelen et al., 2006; Atlas et al., 2015; Timmermans et al., 2015; Peevey and English, 2018). The analysis and forecast errors can be explicitly calculated against the NR, enabling the study of observation impacts, data assimilation performance and short-term forecast errors without concerns regarding the accuracy of verification fields.

As every aspect of an OSSE is simulated, it is necessary to validate the behaviour of important aspects of the OSSE framework against the real world to ensure that experimental results are actually affirmative of real world behaviours. When designing an OSSE framework for evaluating NWP, the goal is to develop a system where the metrics of the simulated system are indistinguishable from similar metrics from the real world. For example, the spatiotemporal distribution of synthetic observations should mimic the distribution of real observations, forecasts should be as skilful as real world operational forecasts, and the impacts of synthetic data should be equivalent to the corresponding impacts of real observations. In practice, only a select subset of such metrics can

^{*}Corresponding author. e-mail: nprive@alum.mit.edu

be matched in both the OSSE and the real world, due to deficiencies in the Nature Run, model error, observation operators and other aspects of the OSSE framework.

The process of adjusting and validating the OSSE against the real world is sometimes referred to as calibration, and often takes the form of an iterative set of steps. One of the most versatile and straightforward methods for calibrating an OSSE is to adjust the simulated observation errors. Other methods such as changing the Nature Run, or the physics or dynamics of the forward NWP model are considerably more challenging and may have unintended side effects. By adjusting the magnitude and characteristics of the observation error, metrics such as the count of ingested observations passing quality control, observation innovation and analysis increment can be calibrated. Some other metrics such as observation impact and forecast skill are also affected somewhat by observation error (Privé et al., 2013).

Observation errors are simulated and added to the synthetic observations because it is generally expected that there is insufficient error introduced to the synthetic observations through the simulation process. In practice, observation errors are considered to include not just flaws in the instrument or calibration, but also representativeness error and observation operator error (Ide et al., 1997). The DAS fundamentally operates by weighing the relative expected error of the observations against the estimated background error, so recreating a realistic juxtaposition of errors in the OSSE framework is crucial. In the real world, these errors are rarely fully understood, and it can be difficult to distinguish errors present in the background field from observation errors. As a result, the simulation of observation errors for an OSSE retains an element of ‘art’ and is not purely an exercise of rote calculation.

Data assimilation systems are designed specifically to account for the presence of random, uncorrelated errors that may be present in the observations. The GSI version for this study uses observation error covariance matrices R that are diagonal, so the DAS performance is optimised for observations that have uncorrelated errors. If the simulated observations have primarily random, uncorrelated errors, the behaviour of the DAS in the OSSE will be closer to optimal than might be expected for real observations that have correlated errors. For portions of the simulated observation errors that are correlated, however, our DAS will also be much less effective in filtering those errors compared with uncorrelated components, and thus a greater portion of the total simulated observation errors will be retained. In practice, only a few published OSSEs include the use of simulated correlated observation errors, for example Halliwell et al. (2014), Privé et al. (2014) and Kleist and Ide (2015). Far more

OSSEs are performed with random uncorrelated observation errors only (ex. Lahoz et al., 2005; Stoffelen et al., 2006; Chen et al., 2011; Aksoy and Lørsolo, 2012; Zhu et al., 2012; Garand et al., 2013; Dutta et al., 2015) or even no simulated observation errors at all (ex. Vecchi and Harrison, 2007; Riishojgaard et al., 2012; Peevey and English, 2018).

The goal of this manuscript is to compare the effects of using solely uncorrelated simulated observation errors in an OSSE to the use of a combination of correlated and uncorrelated errors. The results of this comparison are also validated against NWP with real data to demonstrate the use of correlated observation errors in calibrating an OSSE. The National Aeronautics and Space Administration Global Modelling and Assimilation Office (NASA/GMAO) OSSE framework for NWP (Errico et al., 2013) is used for these experiments, employing the Gridpoint Statistical Interpolation (GSI) DAS (Kleist et al., 2009) and the Global Earth Observing System (GEOS) forecast model (Rienecker et al., 2008).

Forecast skill is often calculated for OSSE experiments, however, these skills are strongly dependent on the model error, particularly for forecasts beyond 24–48 hours. The improvement of the model state due to ingestion of any particular observation type as a fraction of the total error is often greatest at the analysis time and during the initial forecast period. Only a small fraction of initial condition errors will tend to grow with forecast time, the majority of initial condition errors are damped out or overwhelmed by model error growth. Previous work has shown that there are minimal differences in forecast skill after the early forecast period when synthetic observations are used with no added simulated error in comparison to the same observations with a mix of correlated and uncorrelated simulated observation errors (Privé and Errico, 2013). The difference between forecast skill with no simulated observation error and with both correlated and uncorrelated error is expected to be larger than the difference between simulating only uncorrelated error and a combination of correlated and uncorrelated error.

2. Method

The GMAO OSSE has been developed using rigorous validation and careful attention to detail. The Nature Run used is a two-year integration of the GEOS model at approximately 7 km resolution with 72 vertical levels and 30 minute temporal output (Gelaro et al., 2015); this is commonly referred to as the ‘G5NR’ within the community. The 3D-variational GSI DAS along with the GEOS forecast model version 5.17 are used to ingest the synthetic observations and produce forecasts. The version

Table 1. List of simulated correlated errors added to each synthetic observation type.

Observation	Error correlation			
	Horizontal	Vertical	Channel	None
AIRS			X	
AMSU-A	X			
ATMS	X			
CrIS			X	
HIRS4	X			
IASI			X	
MHS	X			
SSMIS	X			
Surface conventional				X
AMV	X	X		
Aircraft				X
Scatterometer				X
RAOB		X		
GPSRO		X		

X denotes the addition of simulated errors with the indicated correlation type.

of the GEOS model used to generate the NR differs from the version used for NWP forecasts in the OSSE in terms of horizontal resolution (7 km vs 25 km), model version, and some choices of model physics such as single versus two-moment cloud microphysics (Barahona et al., 2014) and the boundary layer parametrization. When multiple options for the model physics were available for the GEOS 5.17, the selected physics package was chosen to be different than that used by the G5NR in order to increase model error. This setup is considered a ‘fraternal twin’, meaning that the differences between the Nature Run model and forecast model are smaller than the differences between the forecast model and the real world. However this is not expected to affect the conclusions drawn from this work.

Observations are simulated by a two-step process. The first step applies forward observation operators to the NR fields analogously to the way such operators are applied to the background fields in the DAS. As structures of the NR and background fields differ temporally and spatially and, for some observation types, the observation operator algorithms may differ, the observations produced by this step already include a portion of representativeness error (Errico et al., 2013). The second step adds random errors to the values produced by the first step. These added errors are intended to simulate instrument and additional representativeness errors otherwise missing after the first step. These steps, the observation operators, the error functions, and the tuning of those functions’ parameters are described fully in Errico et al. (2017).

The synthetic observations are chosen to represent the global observing network from July 2015. The data types simulated include conventional data types such as aircraft, rawinsondes, and surface stations; also remote sensing data such as atmospheric motion vectors (AMVs), scatterometers, AMSU-A, MHS, GPS-RO, ATMS, AIRS, CrIS, IASI, HIRS4 and SSMIS. For observations such as aircraft and surface data, the spatiotemporal distribution of simulated observations are based on the corresponding real data from July 2015, with observations determined by interpolating the NR fields. Rawinsondes and dropsondes are likewise simulated to have launch times and locations determined by real observations, but are then advected using the NR wind field. AMVs are simulated based on the NR fields of clouds and humidity (Errico et al., 2020), with the statistics of counts and distribution of observations matched as closely to real observations as possible. GPS-RO data are calculated at the location of real observations, using the ROPP (Culverwell et al., 2015) 2-dimensional operator to simulate the bending angle. Microwave and infra-red radiances are simulated using the Community Radiative Transfer Model (CRTM; Han et al., 2006) and based on the swath footprint from real data, but with cloud contamination using the cloud field from the NR. Full details of the GMAO OSSE framework are described in Errico et al. (2017).

2.1. Simulated errors

For all the observation types, the added simulated errors are random values that may include spatially correlated and uncorrelated added contributions. Both contributions are produced by utilising sets of random values drawn from truncated Gaussian distributions. For the uncorrelated contributions, these are simply independent values scaled according to the tuning procedure to be subsequently described. For the correlated contributions, scaled random values are used to specify random coefficients of principal components of either tuned vertical or channel covariance matrices. Horizontal correlations are created by drawing from random horizontal fields produced using a spectral-transform technique. All these generation algorithms are described mathematically in Errico et al. (2017) along with some examples of principal components and random fields. Table 1 lists the types of correlated errors added to each data type for the simulated observations used in this study. Choices of which types of correlated error to add to each data type were made based on how different the observation (O-F) correlations were in the OSSE and Real data sets. For example, AMSUA showed little missing inter-channel correlation for the active channels, while the hyperspectral instruments showed substantial missing inter-channel

correlation. Similarly, AIRS showed little differences in horizontal correlation of O-F among the thinned observation data set that is actually implemented by the DAS.

The added errors are tuned to better match variances, and spatial or channel correlations of O-F values for each separate observation type or satellite platform (e.g. errors added for the AMSU-A instrument on the NOAA-18 and NOAA-19 satellites are generated independently). As the simulated observations include a certain amount of intrinsic error, only a fraction of the total observation error needs to be simulated and applied to the observations. This is estimated by comparing O-F statistics from a DAS applied to real observations with corresponding statistics from an OSSE previously conducted using the same sets of simulated observation types. By comparing the real and simulated data results, the various parameters required to specify all the error generation functions can be determined. These include variances of correlated and uncorrelated portions of the desired errors to be added, length scales for their spatial correlations, and inter-channel correlations. Examples of the success of this approach appear in Errico et al. (2017) and in Section 3 below.

The tuning algorithm is motivated by an assumption that mismatches between real DAS and OSSE innovation statistics are due to mismatches of observation error characteristics. In the early OSSE development at the GMAO this appeared to be the case and indeed accounts for a large contribution to the discrepancies of innovation statistics obtained without the addition of simulated observation error. It has also become apparent, however, that significant portions of those discrepancies are due to mismatches in statistics of the background error; e.g., if the background error in an OSSE DAS is generally weaker than its real DAS counterpart, then with the same observation error characteristics, the contribution to innovation statistics by errors in background (i.e. errors in the results of the observation operators due to the application of erroneous backgrounds) will also be weaker in the OSSE. When the observation error statistics are modified so that the innovation statistics match under such a circumstance, then these errors will be inflated compared to real ones to compensate for what is missing in the background error. It is difficult to quantitatively distinguish what is realistic from what is over compensation without reliable independent quantitative estimates of real observation error (including instrument, representativeness and observation operator error). When real observation error magnitudes and characteristics are unknown, the ability to interpret the realism of the behaviour of the simulated observation errors in the OSSE is limited.

2.2. Experiment setup

Two different OSSE configurations were compared against a configuration using real data. In one OSSE configuration, the simulated observation error added to the synthetic observations is uncorrelated (NoCorr), while in the second configuration, the simulated observation error is a mix of correlated and uncorrelated error (Corr) that has been calibrated as described in Section 2.1. The magnitude of the standard deviation of the total simulated observation error explicitly added to each data type is the same in both the NoCorr and Corr observation datasets.

Both OSSE configurations are spun up from the beginning of June in the second year of the NR integration. The OSSE is started from a real analysis that differs greatly from the corresponding G5NR synoptic states, but the DAS is spun-up toward the latter by ingesting a dense network of simulated, error-free rawinsonde observations covering the entire globe for a week to adjust the analysis state towards the NR. Starting on 10 June, the error-added simulated observations from the Corr framework are ingested in order to allow the analysis to approach a more realistic state with appropriate error characteristics. The Corr configuration is continued through the end of July, with the 21 days in June treated as a spin-up period for the Corr experiment. The NoCorr configuration begins on 21 June and starts from a background taken from the Corr case. The NoCorr configuration is run through the end of July, again with June treated as a spin-up period.

For the real data configuration (Real), the same DAS and forward model employed in the OSSE are used with real data from 2015. The Real case begins with a spinup period for 20 days in June 2015, followed by cycling for the full month of July 2015. Because the OSSE synthetic dataset is based on the same observational data used by the Real case, the characteristics of the global observing network should be similar when comparing OSSE results to the Real results. However, the synoptic weather that occurs in the NR during this period is not expected to match the real weather from July 2015. Validation of the G5NR (Gelaro et al., 2015) has found that the large scale circulation and behaviour of the G5NR is generally realistic when compared with reanalysis climatologies.

3. Results

First, the degree to which the simulated error affects the correlations of the observation innovations is presented for each data type. Figure 1 shows an example of horizontal correlations of (O-F) for the Real, Corr and NoCorr configurations for AMSU-A NOAA-18 channel 5 calculated globally for the month of July. In the

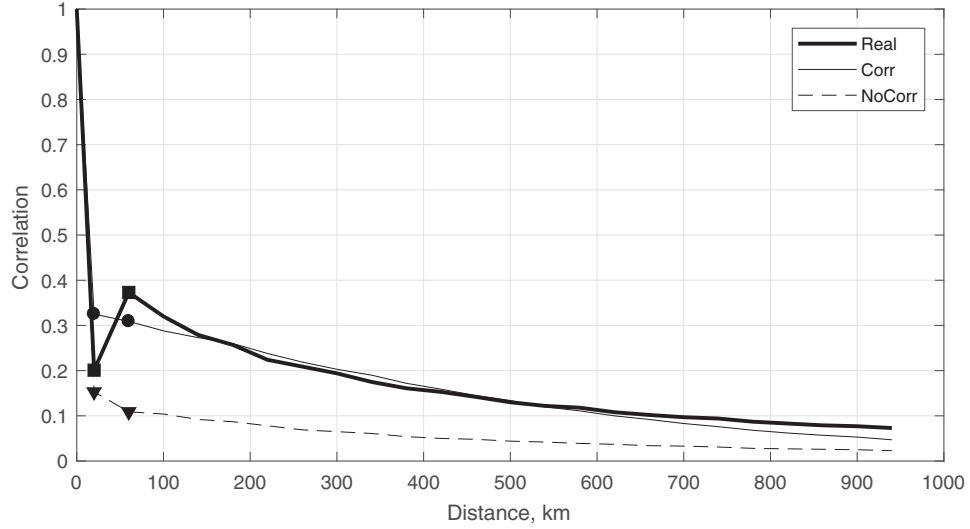


Fig. 1. Horizontal correlation of AMSU-A NOAA-18 channel 5 observation innovations as a function of distance (km) calculated for the month of July. Heavy line, Real case; thin line, Corr case; dashed line, NoCorr case. Markers indicate sample size less than 100 observations per month.

NoCorr case, the horizontal correlation of observation innovations is much smaller than the Real case. For the Corr case, the simulated errors have been added and tuned to match the Real case, and the fit between Corr and Real is very close. Note that the dip in the Real correlation at short distances is due to sampling, as there are few nearby observations due to data thinning. Similarly, [Fig. 2](#) shows an example of vertical correlations for the eastward component of rawinsonde wind observations innovations against innovations at the 600 hPa level. While the simulated error correlations are not a perfect match for the real error correlations, there is better agreement compared to the NoCorr observations, especially between 575 and 625 hPa. Note that the correlations in [Fig. 2](#) are not exactly 1.0 at 600 hPa because the calculation bins observations within the range 593–606 hPa. The fit for vertical correlations is not as good as the fit for horizontal correlations seen in [Fig. 1](#), in part because a single vertical length scale was used for the correlations, regardless of the altitude of the observations. For real observations, the characteristic length scale of correlations has altitude dependence.

The inter-channel correlations (O-F) for IASI metop-a are compared in [Fig. 3](#), calculated globally for the month of July. As expected, the inter-channel correlations for the NoCorr case are relatively weak, especially for the water vapour (channels 294–465) and 10–12 μm window channels (466–616), where correlations in the NoCorr case are less than 30% of correlations in the Real case. The lower left panel in [Fig. 3](#) illustrates the inter-channel correlation of the simulated correlated errors added in the Corr case. In contrast to the NoCorr case, the Corr

case has inter-channel correlations within 20% of the corresponding Real correlations for the channels with high correlations.

By design of the calibration process, the standard deviations of observation innovations for the two cases should be similar. This is demonstrated in the top panels of [Fig. 4](#), where the Corr and NoCorr observation innovations (O-F) are compared with Real for AMSU-A and rawinsonde (RAOB) temperatures calculated globally for the month of July. The standard deviations of O-F are nearly identical for Corr and NoCorr, with overall good agreement with the Real case.

The lower panels of [Fig. 4](#) compare the added simulated error magnitudes to the observation errors assumed by the DAS. For AMSU-A, the added simulated errors have smaller magnitude than the observation errors assigned by the GSI, although for some channels, the simulated error have nearly the same magnitude of the assumed error. However, for rawinsondes in the mid-troposphere, the added simulated errors are greater than the errors assumed by the DAS. This may indicate a region where the errors are overfitted, that is, the added simulated errors are overcompensating for the lack of background error.

3.1. DAS performance

The analysis increment is compared in [Fig. 5](#) between the three configurations for increments of temperature, zonal wind and specific humidity. The Real case has the largest standard deviation of analysis increments, with the Corr case having increments that are approximately 30%

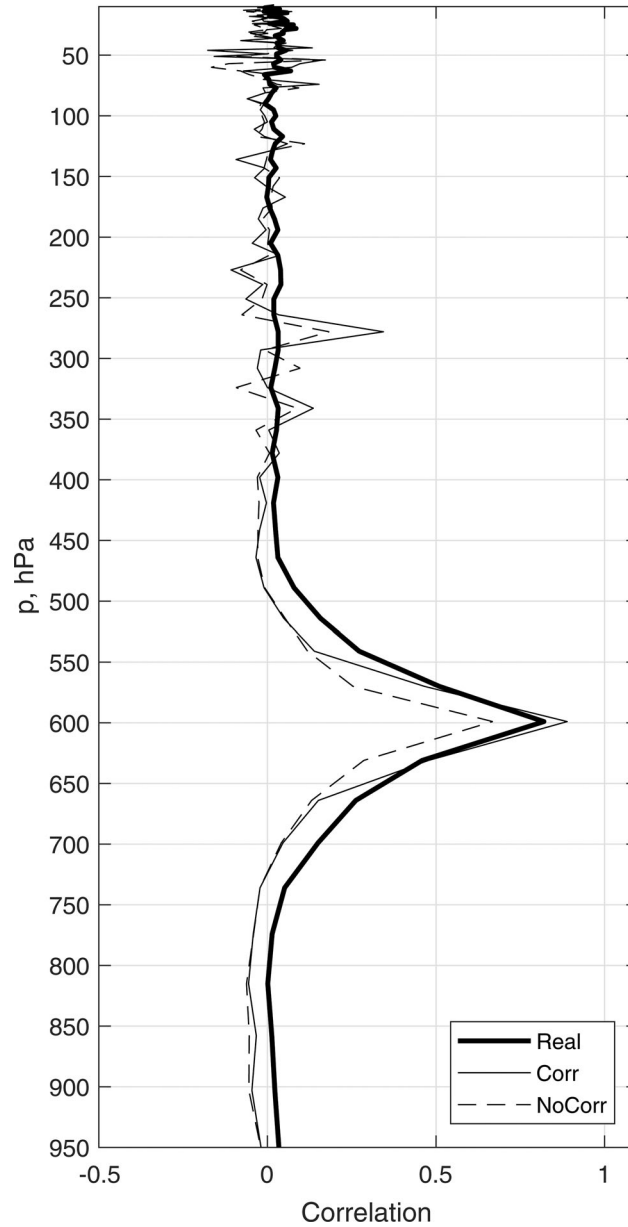


Fig. 2. Vertical correlation of rawinsonde temperature observation innovations as a function of pressure (hPa) for correlations against innovations at 600 hPa. Twice daily data for the month of July. Heavy line, Real case; thin line, Corr case; dashed line, NoCorr case.

smaller than Real, and NoCorr having increments that are 40% smaller than Real. The OSSE analysis increments are smaller than the Real increment due in part to the deficiency of model error growth in the fraternal twin OSSE framework. It is also expected that the model biases are smaller in the OSSE compared to the real world. The forecast model is more similar to the NR model than the forecast model is to the real atmosphere, a common finding in OSSEs. Because the model error growth is weaker in the OSSE, there is less error for the

observations to correct during each DAS cycle, and the analysis increment is therefore smaller.

There are two main mechanisms by which the inclusion of correlated error in the Corr case results in larger analysis increments compared to the NoCorr case. First, the DAS is designed to handle uncorrelated observation errors but does not consider correlated errors, thus the performance of the DAS is suboptimal when observations include correlated errors. The analysis increment includes not just useful information content from the observations

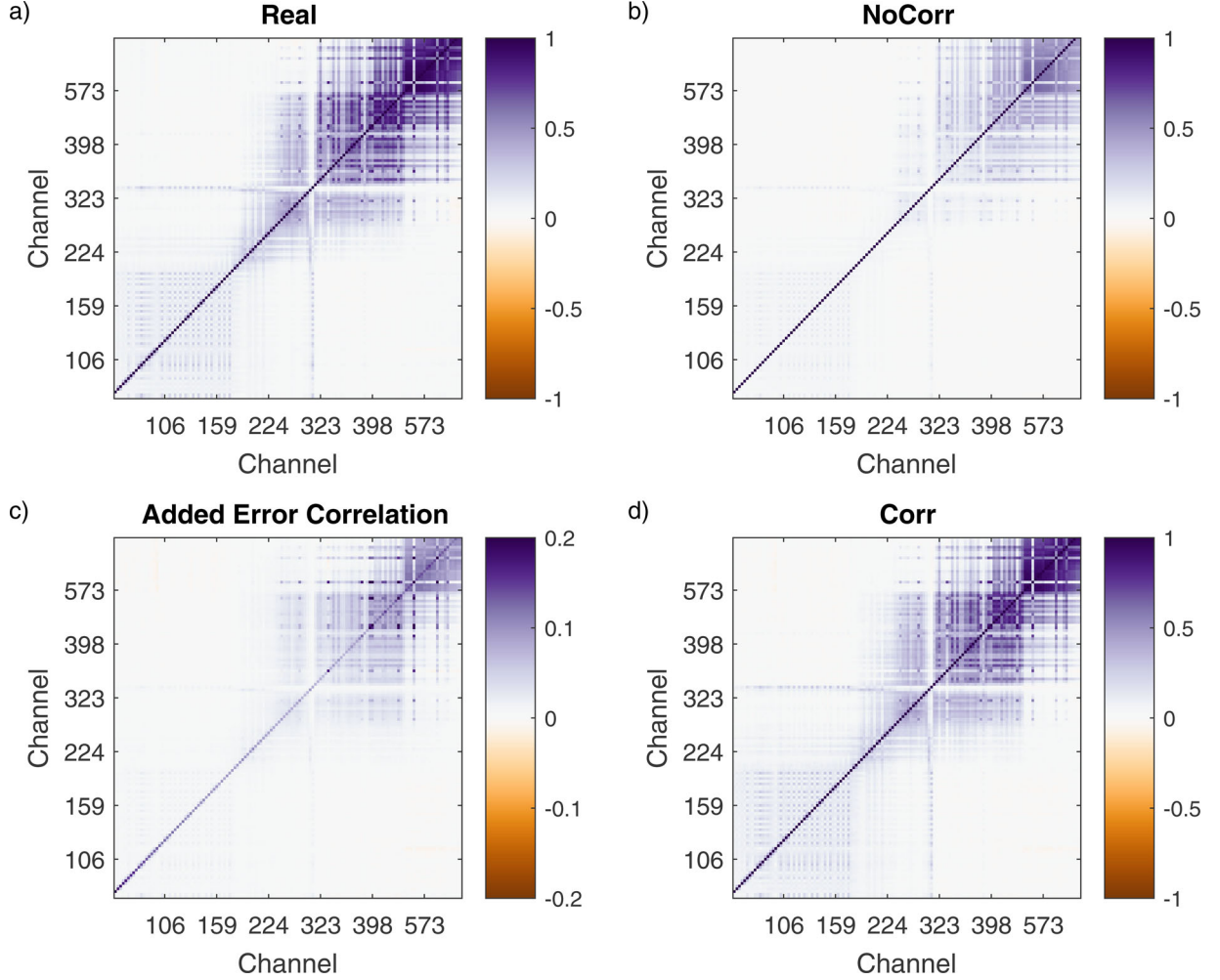


Fig. 3. Channel correlations for IASI metop-a observation innovations. Twice daily data for the month of July. (a) Real case (O-F); (b) NoCorr case (O-F); (c) simulated error correlations added for Corr case; (d) Corr case (O-F).

but also observation error, both of which are spread geographically by the DAS. While the DAS is very effective at filtering much of the error in the NoCorr case, by design it is less effective at filtering the correlated error in the Corr case, so a greater portion of the error is incorporated into the analysis increment.

The second reason for larger analysis increments in the Corr case than the NoCorr case is the relative quality of the background states. As the truth is available in the OSSE in the form of the Nature Run, the analysis and background errors can be explicitly calculated by verifying against the NR fields. Figure 6 shows the RMS background error for the temperature, wind, and specific humidity in the Corr case (left), and the difference in background error between the NoCorr and Corr cases (right). The RMS background error in the NoCorr case is as much as 5–10% lower in magnitude than for the Corr case in some areas. As a result of the smaller background

error, there is less work for the observations to perform in the NoCorr case.

One useful metric is the difference between the absolute analysis and background errors, $|A - NR| - |B - NR|$, where A and B are the analysis and background fields respectively and NR is the Nature Run interpolated to the same resolution as A and B . This metric quantifies where the analysis is improved (or degraded) compared to the background, with negative values indicating improvement. Figure 7 illustrates this metric for the Corr case and the difference between the NoCorr and Corr cases for temperature, specific humidity, and zonal wind.

The largest improvements due to the analysis increment are seen in the tropics, a region where model error is relatively large and where errors grow rapidly due to moist physical processes. In the extratropics, the performance of the DAS is more mixed, with global improvements in specific humidity. For some areas such as the

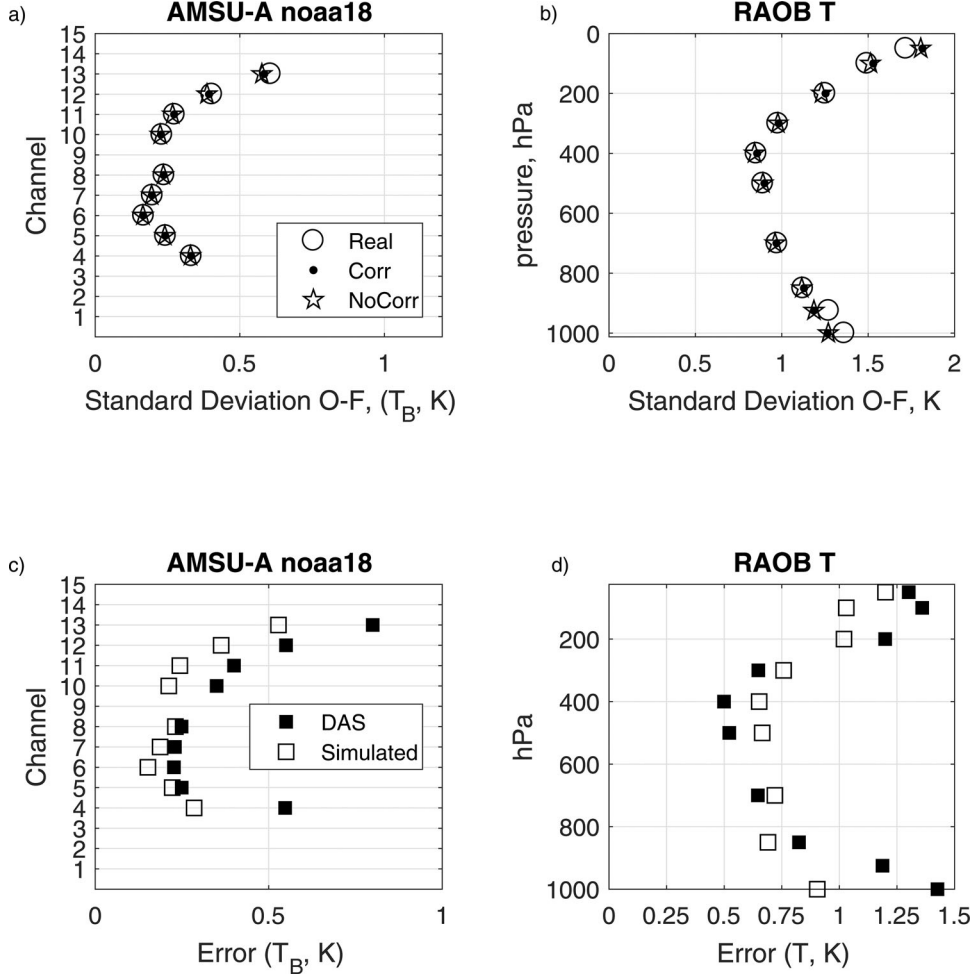


Fig. 4. Top, Standard deviations of observation innovations (O-F). Twice daily data for the month of July. Real case, open circles; Corr case, solid dots; NoCorr case, open stars. Bottom, comparison of the magnitude of simulated added errors (open squares) and the R observation error weighting used by the DAS (dark squares). (a, c) AMSU-A noaa-18 as a function of channel; (b, d) rawinsonde temperature as a function of height.

polar regions and near the tropopause the net effect of the ingestion of observations is detrimental to the analysis for temperature and winds. The presence of correlated observation errors in the Corr case has an overall tendency to degrade the performance of the DAS, with the NoCorr case showing more negative (beneficial) impacts of the DAS compared to the Corr case (right panels in Fig. 7). In some regions of the extratropics, the NoCorr case shows a beneficial contribution from the DAS while the Corr case has a detrimental DAS contribution. Regions where there is a time-mean degradation of the analysis state compared to the background implies suboptimality of the DAS framework.

Similar to the background error, the analysis error for the NoCorr case is approximately 5–10% smaller than the Corr case, as shown in Fig. 8. Although the analysis increment is larger for Corr, the effective number of

independent observations is smaller than for NoCorr, and the initial background state also has greater magnitude of error for the Corr case.

A small area of increased background and analysis error for NoCorr compared to Corr is seen for zonal wind on the equator near 100 hPa. However, Fig. 5 shows little change in the analysis increment for zonal wind in this region, although there is a greater magnitude of RMS temperature increment. Likewise, Fig. 7 shows more beneficial temperature increments at the tropical tropopause for NoCorr, while zonal wind increments are neutral to slightly detrimental at 100 hPa but with little difference between NoCorr and Corr. The zonal mean zonal wind analysis bias (not shown) has an easterly maximum on the equator at the tropopause that may be partially related to suboptimal balance between the wind and temperature fields in the DAS. This easterly wind bias is

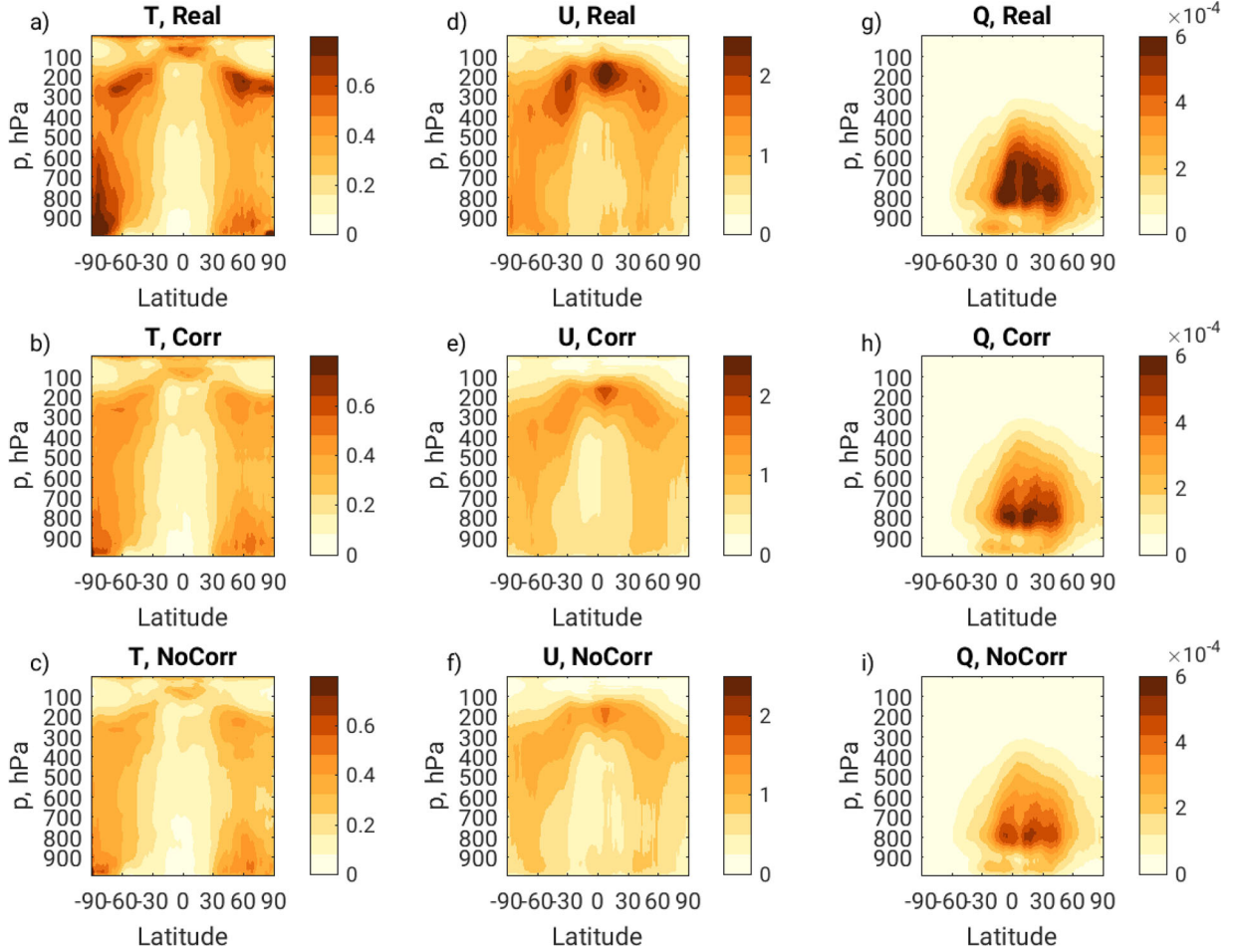


Fig. 5. Zonal means of temporal root mean square analysis increments (A, B). Twice daily data for the month of July. (a, d, g) Real case; (b, e, h) Corr case; (c, f, i) NoCorr case. (a, b, c) T (K); (d, e, f) zonal wind (m s^{-1}); (g, h, i) specific humidity (kg kg^{-1}).

slightly (5%) stronger in the NoCorr case than the Corr case at 100 hPa. It is possible that the addition of correlated errors acts to smooth out features that cause the easterly wind bias.

3.2. Forecast errors

Five-day forecasts are produced daily for the month of July starting at 0000 UTC for both the NoCorr and Corr cases. The fractional difference in forecast RMSE between the NoCorr and Corr cases is shown in Fig. 9 for the NHEx (20N-90N), SHEX (20S-90S), and Tropics (20S-20N) regions as a function of model level pressure, where negative values indicate lower error for the NoCorr case, and stippling indicates significance at the 90% confidence level. As expected, the NoCorr case has smaller forecast errors in the troposphere during the initial forecast period, with greatest differences between Corr and NoCorr occurring at the initial time. For most fields, differences are no

longer significant after 3 days, although the sample size here is relatively small with only 31 forecasts.

The largest fractional improvement in forecast skill is seen for the SHEX region, with approximately 5% improvement in RMS temperature and zonal wind errors during the initial forecast period. This may be due in part to seasonality, as July is winter in the SHEX region. The Tropics region shows the smallest differences in forecast error between Corr and NoCorr, with improvements of approximately 1-2% or less for temperature and wind during the initial forecast period. In the Tropics, the growth of model errors through fast processes such as convection tend to override improvements in the initial conditions during the early forecast period.

3.3. Observation impacts

The GEOS OSSE includes the capability for forecast sensitivity to observations (FSO) using a moist adjoint

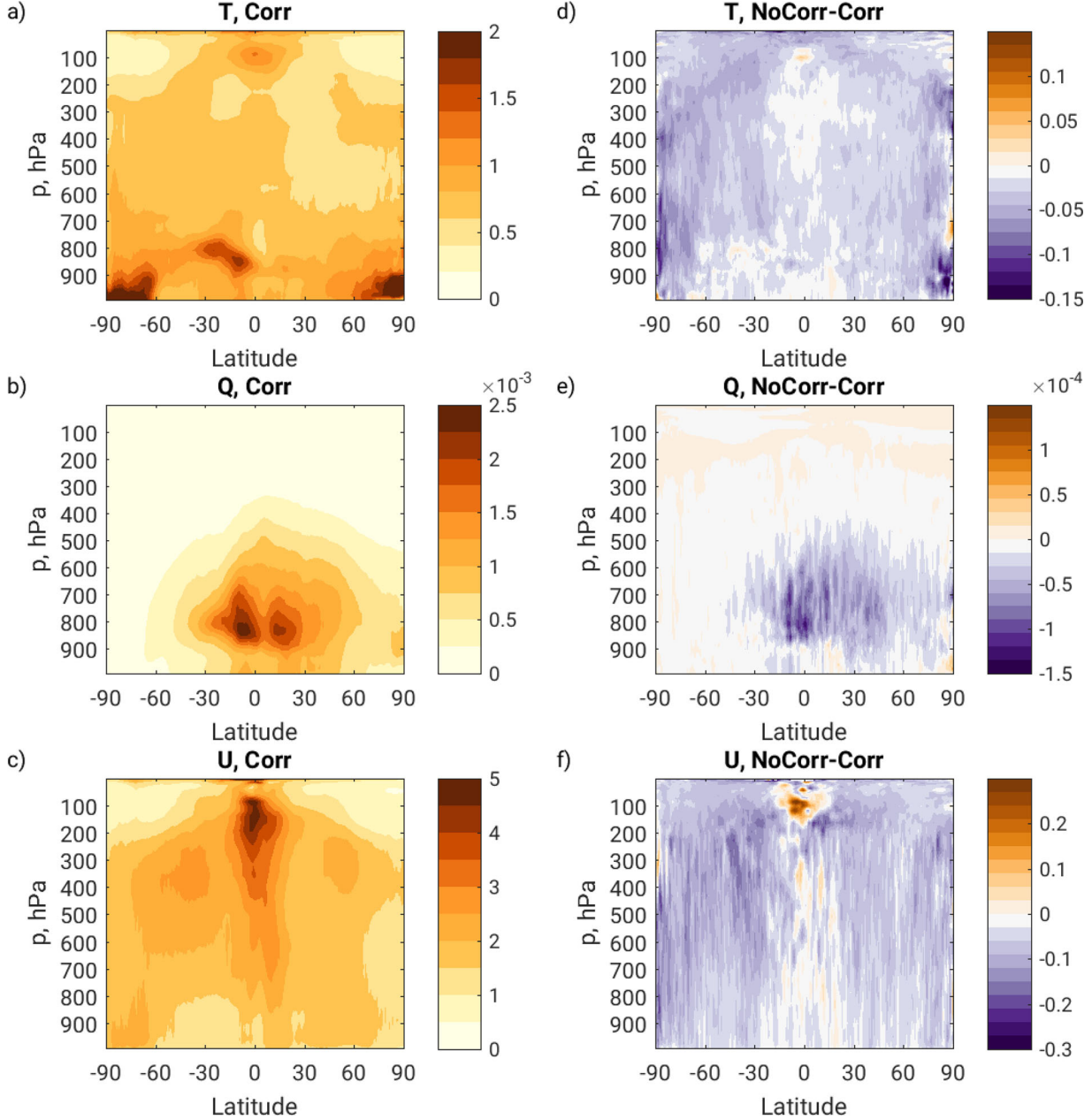


Fig. 6. Zonal means of root temporal mean square background errors. (a, d) Temperature (K); (b, e) specific humidity (kg kg^{-1}); (c, f) zonal wind (m s^{-1}). (a, b, c) Left panels, Corr case; (d, e, f) difference between NoCorr and Corr cases.

described by Holdaway et al. (2014). The total wet energy is selected as the error norm as in Privé and Errico (2019), with self-analysis verification in order to compare the Real and OSSE configurations. The resulting estimates of observation impact are shown in Fig. 10, where negative impacts indicate a decrease in forecast error. The Corr and NoCorr observation impacts are considerably smaller in magnitude than the impacts for the Real case, which is expected due to insufficient model error growth and bias in the OSSE (Privé and Errico, 2019).

The Corr case has slightly more negative (beneficial) impacts for most data types compared to the NoCorr case, with statistically significant differences for AMSU-A, rawinsondes and AMVs. However, IASI, MHS and AIRS have more positive (detrimental) impacts in the Corr case. In general, when larger magnitudes of simulated observation error are added to all data types, greater observation impacts result (Privé et al., 2013). When observation errors are increased for only a subset of data types, other observations may have more

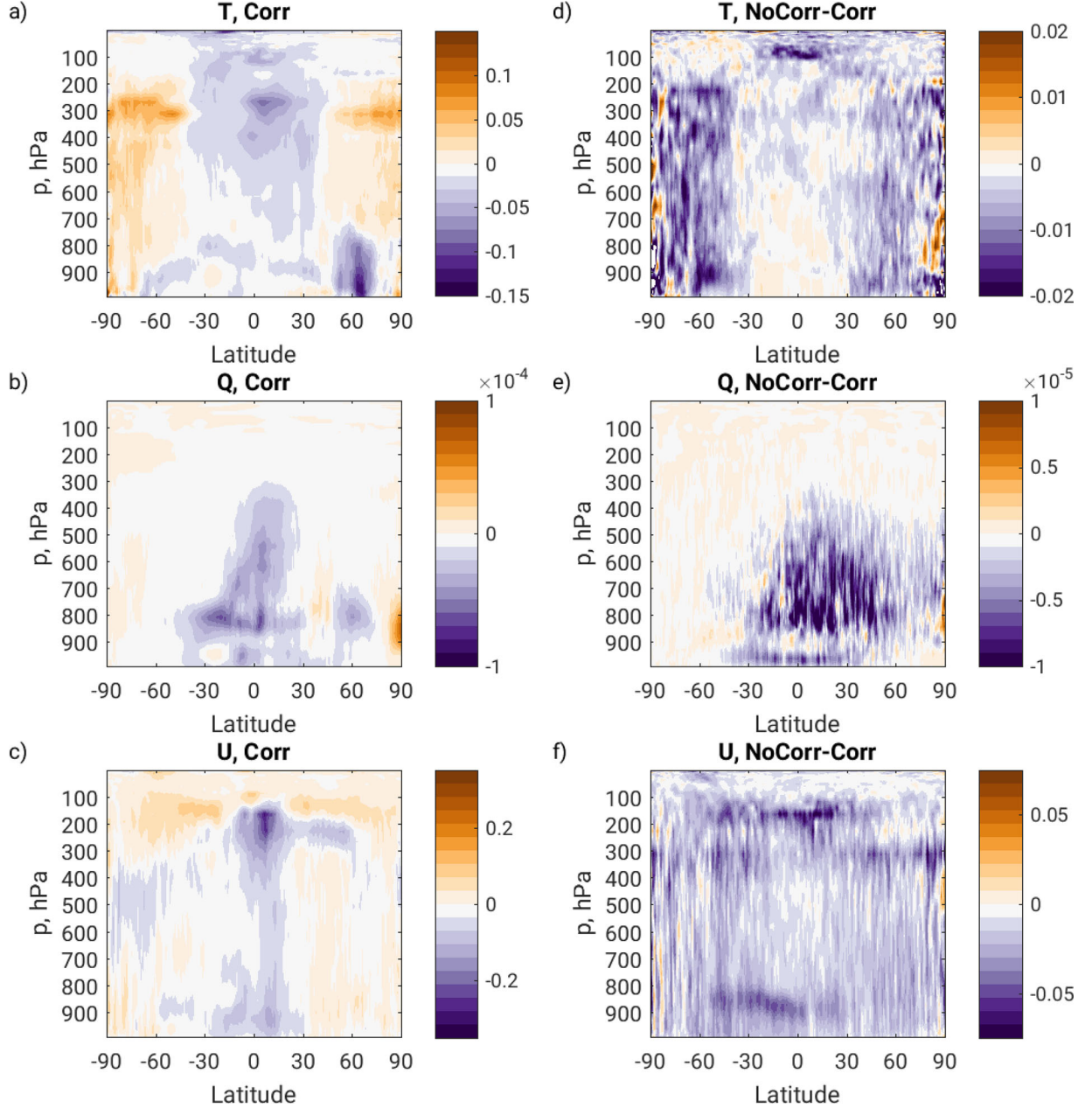


Fig. 7. Zonal means of temporal mean differences in absolute analysis and background errors, $|A-NR| - |B-NR|$. (a, d) Temperature (K); (b, e) specific humidity (kg kg^{-1}); (c, f) zonal wind (m s^{-1}). (a, b, c) Corr case; (d, e, f) difference between NoCorr and Corr cases.

beneficial impacts as these types compensate for the degraded observations.

The tendency for greater observation errors to lead to adjoint estimates of larger beneficial observation impacts can be counterintuitive. FSO is based on the difference in 24-hour forecast skill between pairs of forecasts, one initialised from the background state and the other initialised from the analysis state. The forecast error difference between these forecast pairs is a combination of the growth of model error and initial

condition error. Here, both the background and the analysis states have lower error for the NoCorr case compared for the Corr case. In the Corr case, the greater initial condition error in both the background and analysis states results in greater error growth for both members of each forecast pair, with a slight increase in the difference between the pairs of forecast skills compared to NoCorr. What the FSO does not capture well is the overall degradation of both the analysis and background states due to the introduction

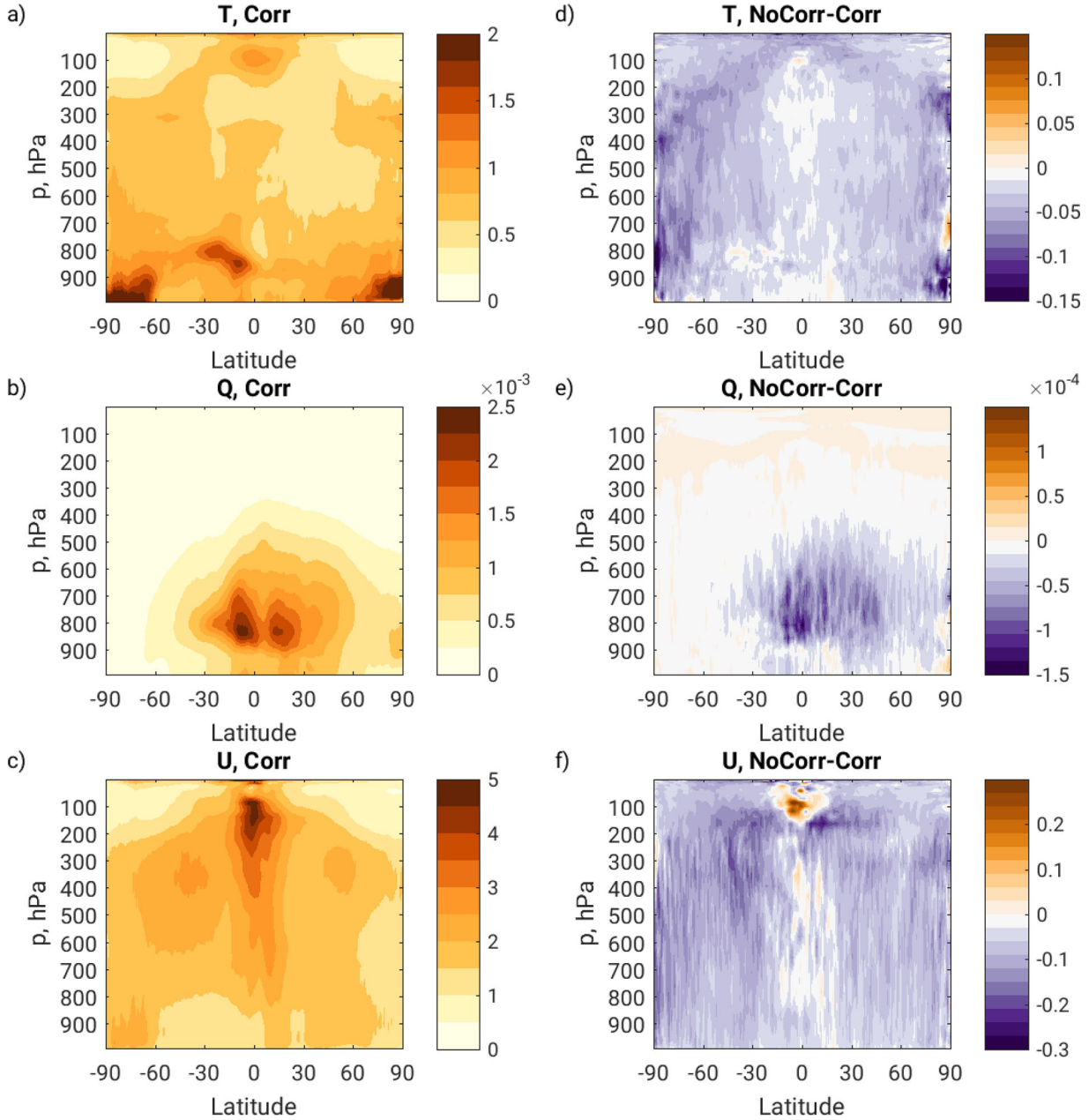


Fig. 8. Zonal means of root temporal mean square analysis errors. (a, d) Temperature (K); (b, e) specific humidity (kg kg^{-1}); (c, f) zonal wind (m s^{-1}). (a, b, c) Corr case; (d, e, f) difference between NoCorr and Corr cases.

of correlated errors in the Corr case, because they are degraded in tandem. In the hypothetical situation of a perfect model with trivially small initial condition error, the observations would have only a tiny impact as calculated by FSO methods.

The differences in observation impacts between the Corr and NoCorr cases imply that the addition of correlated observation errors more strongly affects the ability of the DAS to adequately handle ingestion of

some data types compared to others. Both IASI and AIRS are hyperspectral instruments with channel-correlated errors added in the Corr case, with strong correlations seen between some sets of channels (Fig. 3). AMSU-A, on the other hand, has a strong increase in the magnitude of beneficial impact in the Corr case. The increase in AMSU-A impact may be compensation for the degradation of other observing types, such as AIRS and IASI.

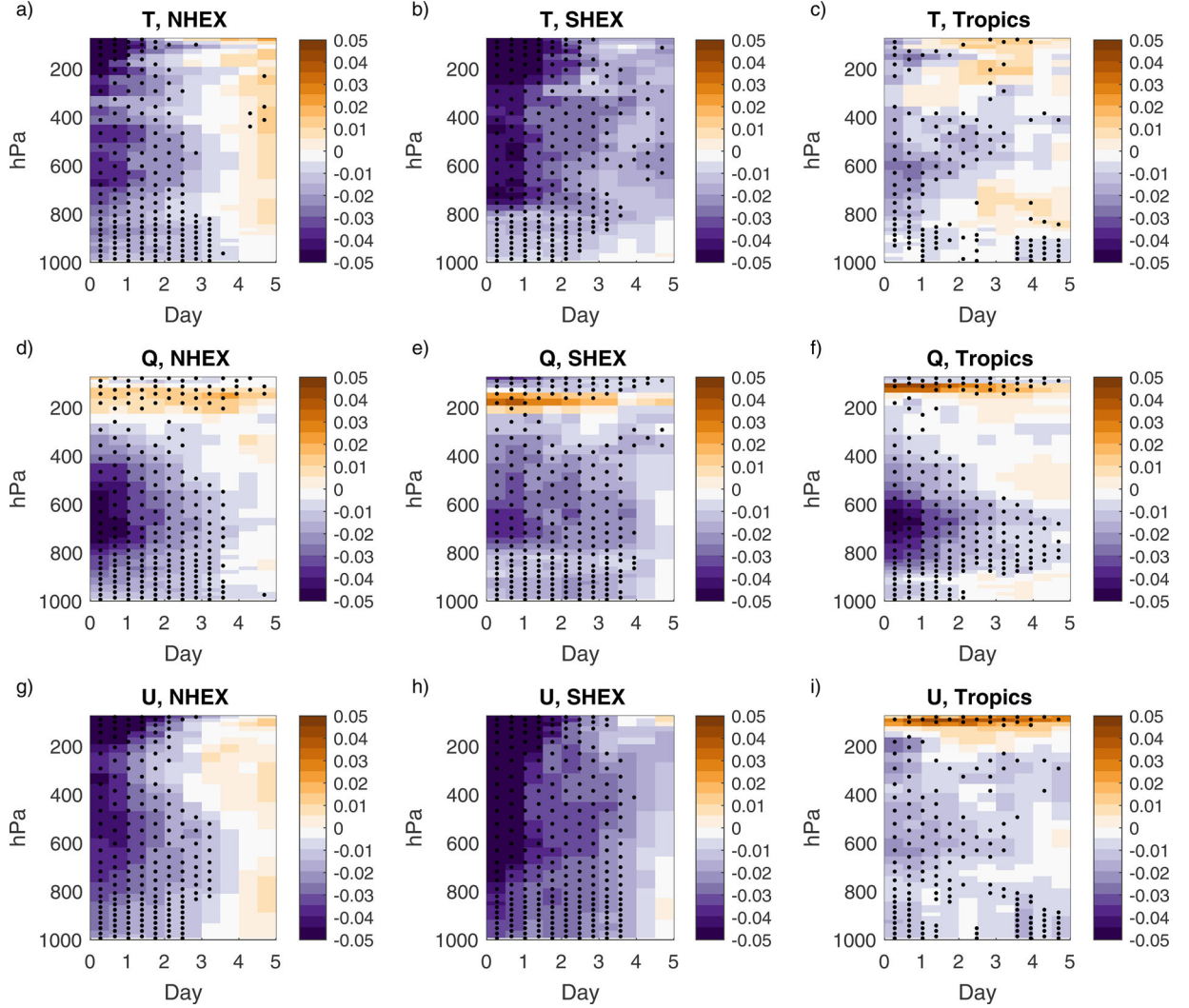


Fig. 9. Fractional difference in areal mean temporal-RMS forecast error, (NoCorr-Corr)/Corr, month of July. Stippling indicates significant difference at the 90th percentile confidence level. (a, b, c), temperature error; (d, e, f) specific humidity error; (g, h, i) zonal wind error. (a, d, g) NHEX (20N-90N); (b, e, h) SHEX (90S-20S); (c, f, i) Tropics (20N-20S).

4. Implications for OSSEs

The results of this study show that the use of correlated observation errors enhances the capability for calibrating the statistics of DAS behaviour in the OSSE to more closely match the real world behaviour. The vast majority of published NWP OSSE studies either apply only uncorrelated simulated observation errors, or omit simulated observation errors entirely. Efforts to calibrate and validate the OSSE performance are likewise not widespread in the published literature.

Simplified approaches to observation error and calibration can be suitable for some idealised or proof-of-concept studies. However, for OSSEs that are intended to put a new observing system in context with the existing global observation network, or to explore the behaviour

of DAS processes, it is important to try to replicate real-world behaviour of the entire NWP system as faithfully as possible in the OSSE. Calibration and validation are essential to ensuring the robustness of OSSE results and to understanding the limitations of OSSE capabilities. The inclusion of correlated observation errors both increases the realism of the synthetic observations and how these observations are ingested by the DAS, and also allows an additional avenue for calibration of the OSSE framework.

For the GMAO OSSE framework, the statistics of observation innovation have been chosen as one of the primary metrics for calibration. The calibration proceeds by first adjusting the uncorrelated observation errors to match the standard deviation of observation innovation

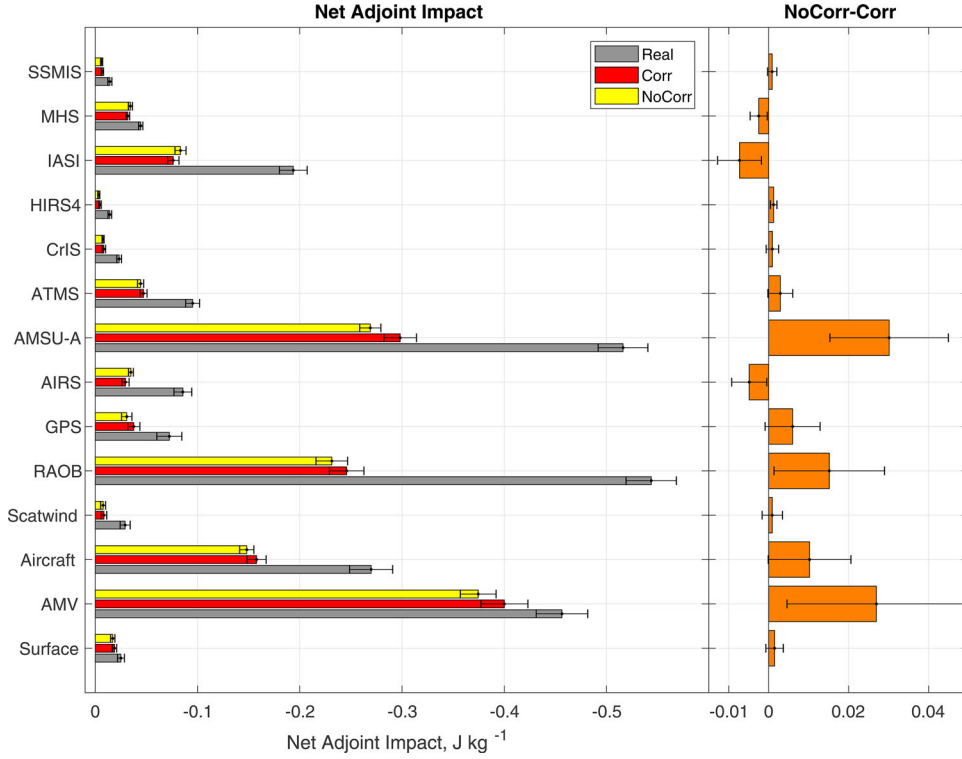


Fig. 10. Daily mean FSO estimates of observation impacts on 24-hour forecast skill for a total wet energy norm, self-analysis verification. Left panel, comparison of net daily impact with whiskers indicating confidence interval at the 95th percentile; right panel, difference in impacts between NoCorr and Corr cases, with whiskers indicating the confidence interval of paired differences at the 95th percentile.

in the OSSE for each data type to that calculated for real data in the same model/DAS system. Then the simulated errors are partitioned between uncorrelated and correlated error to additionally match the correlations of observation innovations for real data. The NoCorr case shows that when calibration with observation innovation is performed with uncorrelated simulated errors only, the match between the real and OSSE cases is poorer for statistics of analysis increments compared to the match between the real and Corr cases.

Tuning of the Corr case simulated errors involves adjusting the observation error contribution to O-F covariances. However, if the discrepancy between the real and OSSE O-F covariances is largely due to incorrect correlations in the background field, then adding correlated error to the observations can only correct for a portion of the missing covariance of observation innovation. In particular, if independent errors are added to distinct data types, correlations between O-F for two different data types that could result from correlation of the background field will be absent.

If a different metric were chosen for calibration such as analysis increment or observation impact, it is anticipated that much larger magnitude of observation errors

would be required for calibration. Calibration via observation error is limited in that there is a lower threshold of calibration where no error is added to the simulated observations, and an upper limit at which the quality control from the DAS begins to remove an excessive number of observations.

One caveat when calibrating the synthetic observations is that most NWP OSSEs have weaker model error than is seen for operational NWP models in the real world. As seen in Figs. 5 and 10, this insufficient model error strongly affects the analysis increment and observation impacts, as there is less “work” for the observations to perform in correcting model error growth. The deficiency in model error is also expected to affect the background and analysis states, although these cannot be calculated directly with real data since the true state of the atmosphere is not known. This will in turn affect metrics such as the observation innovation, as the background state is expected to be too accurate in the OSSE in comparison to a real world NWP system. Thus, by matching the statistics of (O-F) during calibration, the observations artificially take on a portion of error that is associated with the background state in the real world. This overcompensation for the lack of model error by degrading the

quality of the observations becomes more substantial for metrics that are even more strongly affected by the model error, such as the analysis increment or FSO. The dependency on model error growth should be considered when selecting which metrics to use when calibrating an OSSE.

It is notable that although including correlated observation error in the simulated observations results in a better calibration of the analysis increment, the analysis increments in the Corr case are still substantially smaller than for Real. Some of this discrepancy may be due to the insufficiency of model error in the fraternal twin OSSE framework. Attempting to increase the model error is difficult, as ideally the NR would be generated using the most realistic model available, and the option of degrading the performance of the forecast model would mean that the experiment results are not representative of the behaviour of modern operational systems. The deficiencies of NWP models compared to the real world are only partially understood, and it would be preferred to attempt to achieve these same deficiencies in the OSSE framework. For the GMAO OSSE, attempts to increase the model error by using a different convective microphysics scheme and boundary layer parameterisation resulted in only very small changes in the performance of the OSSE system. Previous experiments using the T511 ECMWF NR in the GMAO OSSE also indicated a substantial lack of model error with much more disparate models. It is thus not clear how model error could be increased sufficiently without sacrificing realism in other aspects of the OSSE setup.

When choosing to calibrate the OSSE framework through the addition of simulated observation errors rather than by manipulation of model error, the application of realistic types of observation error is particularly important. The simulated correlations do not include temporal correlations, which are likely to be present in at least some real-world observing types. Correlations between different data types that observe the atmosphere in the same spatiotemporal region are also neglected, but these could be significant if the background state in the OSSE has weaker variance than a real world background state. The attempt here to include spatially correlated errors is merely a first step towards simulating realistic observation errors for an OSSE. The ability to include more realistic and sophisticated simulated errors is limited by the current understanding of errors for real observations. The simple simulated error correlations introduced here have a modest impact on the performance of the OSSE framework. It is a reasonable assumption that achieving a highly realistic OSSE will require many incremental improvements in the modelling of observation errors.

In the OSSE experiments, both model and observation biases are expected to be considerably smaller in magnitude than in the real world. In the GMAO OSSE, there was no effort to simulate poorly understood biases that are not generally removed by the bias correction algorithms due to the difficulty in simulating these biases. However, biases are introduced through the bias correction routines for the DAS that act to adjust some types of observations to remove perceived bias. In actuality, these bias correction routines may correctly remove some observation biases, but may also introduce biases when model error is incorrectly ascribed to observations. In the GMAO OSSE, the bias correction is allowed to make these adjustments, but the majority of adjustments are expected to be due to model error and not to the smaller biases introduced through operator error and such. Biases in observation types that are not subject to bias correction, such as most conventional types, might be expected to more strongly impact the analysis quality and observation impact. For example, known biases for GPS-RO bending angles due to operator error that occurs for real observations also act on the simulated GPS-RO observations and are the chief culprit in the degraded temperature field at 300 hPa seen in Fig. 7a. As further research on observation biases leads to better understanding of the sources and magnitudes of these biases, it may be possible to simulate observation biases in a realistic way. Necker et al. (2018) and Kotsuki et al. (2019) have found that biases in the forecast model and in the observations can have a large effect on FSO calculations. Some aspects of the effect of missing biases in an OSSE are discussed in Privé et al. (2021).

The greatest impact of correlated observation errors is seen in the early forecast period, with small or neutral impacts beyond 72 hours. However, when performing OSSEs for new instruments, the greatest impact of these instruments is also generally seen during this initial forecast period (ex., McCarty et al., 2021). Incorporating correlated simulated observation errors as in the Corr case might then have some effect on the impacts of proposed new instruments on the short-term forecast skill.

As seen in the FSO results, some instrument types are more strongly affected by the addition of correlated observation errors than others, with differences between the Corr and NoCorr impacts of up to 10% for select types. This may be in part due to the magnitude of correlated error, which differs between each observation type, but may also be related to the ability of the DAS to handle correlated errors for that particular data type. When introducing a new proposed observation type for which real data are not available, some consideration should be given to what types of correlated error might be expected for the data type.

There are many variables that can affect the characteristics of observation error, including regional or altitude dependence, effects of various forms of moisture, or errors that may vary with synoptic condition, for just a few examples. These influences may result in different error magnitudes and types of observation error correlations. When performing an OSSE on a new observing type that does not have existing observations for calibration, it is especially important to consider both how simulated errors are added to the synthetic observations, and how the DAS handles the new observation type. One option is to test the new observation type with a range of different error magnitudes and types in order to test the robustness of the OSSE results. For example, the new observations could first be tested with no added simulated observation errors and strong weighting in the DAS in order to find a range of maximum impact. Then simulated errors could be added to the new observation type and ingested by a DAS with less weighting to find a “break-even” point at which the new observations have minimal beneficial impact on the analyses or forecasts. This type of careful experimentation can provide useful information to both the instrument development and DAS development efforts that can facilitate the eventual ingestion of real observations.

A limitation of this study is the use of 3D-Variational data assimilation, while most current operational NWP models use some form of 4D-Variational (4DVar) data assimilation. With 4DVar DAS, the background error covariances include a dependency on the unique synoptic situation for each cycle time. The GMAO OSSE has more recently been implemented using a hybrid 4D-Ensemble Variational DAS with the same global circulation model as in these experiments. The relative performance of the OSSE compared to real observations in the 4DVar framework is similar to that seen with 3DVar. For example, the analysis increments in the OSSE are approximately the same fraction of increments with real observations, and the FSO estimations of observation impacts are also very close to that seen with 3DVar. This similar behaviour between different DAS versions is encouraging that the results from this study may hold with the 4DVar framework.

Acknowledgements

Support for this project was encouraged by Steven Pawson and funded by NASA/GMAO core provided under the Modeling, Analysis, and Prediction Program. 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. 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. Helpful comments by three anonymous reviewers lead to significant improvements in the manuscript.

Data availability statement

The dataset on which this paper is based is too large to be retained or publicly archived with available resources. Documentation and methods used to support this study are available from Nikki Privé at NASA/GMAO.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

- Aksoy, A. and Lorsolo, S. 2012. The HWRF Hurricane Ensemble Data Assimilation System (HEDAS) for high-resolution data: the impact of airborne doppler radar observations in an OSSE. *Mon. Weather Rev.* **140**, 1843–1862. doi:10.1175/MWR-D-11-00212.1
- Atlas, R., Hoffman, R. N., Ma, Z., Emmitt, G. D., Wood, S. A. and co-authors. 2015. Observing system simulation experiments to evaluate the potential impact of an optical autocovariance wind lidar on numerical weather prediction. *J. Atmos. Ocean Technol.* **32**, 1593–1613. doi:10.1175/JTECH-D-15-0038.1
- Barahona, D., Molod, A., Bacmeister, J., Nenes, A., Gettelman, A. and co-authors. 2014. Development of two-moment cloud microphysics for liquid and ice within the NASA Goddard Earth Observing model GEOS-5. *Geosci. Model Dev.* **7**, 1733–1766. doi:10.5194/gmd-7-1733-2014
- Chen, S.-H., Chen, J., Chang, W., Lin, P., Lin, P. and co-authors. 2011. Observing system simulation experiment: development of the system and preliminary results. *J. Geophys. Res.* **116**.doi:10.1029/2010JD015103
- Culverwell, I., Lewis, H., Offiler, D., Marquardt, C. and Burrows, C. 2015. The radio occultation processing package, ROPP. *Atmos. Meas. Tech.* **8**, 1887–1899. doi:10.5194/amt-8-1887-2015
- Dutta, S., Garand, L. and Heilliette, S. 2015. Impact evaluation of assimilating surface sensitive infrared radiance observations over land and sea ice from observing system simulation experiments. *Adv. Meteorol.* **2015**, 1–9. doi:10.1155/2015/847561
- Errico, R., Carvalho, D., Privé, N. and Sienkiewicz, M. 2020. Simulation of atmospheric motion vectors for an observing system simulation experiment. *J. Atmos. Ocean Technol.* **37**, 489–505. doi:10.1175/JTECH-D-19-0079

- Errico, R. and co-authors. 2017. Description of the GMAO OSSE for weather analysis software package: version 3. Technical Report 48, 156 pp., NASA/TM-2017-104696, National Aeronautics and Space Administration.
- Errico, R. M. and Privé, N. C. 2018. Some general and fundamental requirements for designing observing system simulation experiments (OSSEs). *WMO Rep. WWRP 2018-8*, 33 pp. Online at: https://www.wmo.int/pages/prog/arep/wwrp/new/documents/Final_WWRP_2018_8.pdf
- Errico, R. M., Yang, R., Privé, N., Tai, K.-S., Todling, R. and co-authors. 2013. Validation of version one of the observing system simulation experiments at the global modeling and assimilation office. *Q. J. R. Meteorol. Soc.* **139**, 1162–1178. doi:10.1002/qj.2027
- Garand, L., Feng, J., Heillette, S., Rochon, Y. and Trishchenko, A. P. 2013. Assimilation of circumpolar wind vectors derived from highly elliptical orbit imagery: impact assessment based on observing system simulation experiments. *J. Appl. Meteorol. Clim.* **52**, 1891–1908. doi:10.1175/JAMC-D-12-0333.1
- Gelaro, R. and co-authors. 2015. Evaluation of the 7-km GEOS-5 nature run. NASA/TM2014-14606, Technical Report, NASA.
- Halliwell, G. J., Srinivasan, A., Kourafalou, V., Yang, H., Willey, D. and co-authors. 2014. Rigorous evaluation of a fraternal twin ocean OSSE system for the open Gulf of Mexico. *J. Atmos. Ocean Technol.* **31**, 105–130. doi:10.1175/JTECH-D-13-00011.1
- Han, Y., van Delst, P., Liu, Q., Weng, F., Yan, B. and co-authors. 2006. JCSDA Community Radiative Transfer Model (CRTM) - version 1. Technical report, OAA Technical Report 122. pp. 33, NOAA, Silver Spring, MD.
- Holdaway, D., Errico, R., Gelaro, R. and Kim, J. 2014. Inclusion of linearized moist physics in NASA's Goddard Earth Observing System data assimilation tools. *Mon. Weather Rev.* **142**, 414–433. doi:10.1175/MWR-D-13-00193.1
- Ide, K., Courtier, P., Ghil, M. and Lorenc, A. 1997. Unified notation for data assimilation: operational, sequential and variational. *J. Meteorol. Soc. Jpn.* **75**, 181–189. doi:10.2151/jmsj1965.75.1B_181
- Kleist, D. and Ide, K. 2015. An OSSE-based evaluation of hybrid variational-ensemble data assimilation for the NCEP GFS. Part I: system description and 3d-hybrid results. *Mon. Weather Rev.* **143**, 433–451. doi:10.1175/MWR-D-13-00351.1
- Kleist, D., Parrish, D., Derber, J., Treadon, R., Wu, W.-S. and co-authors. 2009. Introduction of the GSI into the NCEP global data assimilation system. *Weather Forecast.* **24**, 1691–1705. doi:10.1175/2009WAF2222201.1
- Kotsuki, S., Kurosawa, K. and Miyoshi, T. 2019. On the properties of ensemble forecast sensitivity to observations. *Q. J. R. Meteorol. Soc.* **145**, 1897–1914. doi:10.1002/qj.3534
- Lahoz, W., Brugge, R., Jackson, D., Migliorini, S., Swinbank, R. and co-authors. 2005. An observing system simulation experiment to evaluate the scientific merit of wind and ozone measurements from the future SWIFT instrument. *Q. J. R. Meteorol. Soc.* **131**, 503–523. doi:10.1256/qj.03.109
- McCarty, W., Carvalho, D., Moradi, I. and Privé, N. C. 2021. Observing system simulation experiments investigating a constellation of 4–5 μm infra-red sounders. *J. Atmos. Ocean Technol.* doi:10.1175/JTECH-D-20-0109.1
- Necker, T., Weissman, M. and Sommer, M. 2018. The importance of appropriate verification metrics for the assessment of observation impact in a convection-permitting modelling system. *Q. J. R. Meteorol. Soc.* **144**, 1667–1680. doi:10.1002/qj.3390
- Peevey, T. R. and English, J. M. Improving winter storm forecasts with observing system simulation experiments. Part I: an idealized case study of three U.S. storms. *Mon. Weather Rev.* **146**, 1341–1366. doi:10.1175/MWR-D-17-0160.1
- Privé, N. and Errico, R. 2019. Uncertainty of observation impact estimation in an adjoint model investigated with an observing system simulation experiment. *Mon. Weather Rev.* **147**, 3191–3204. doi:10.1175/MWR-D-19-0097.1
- Privé, N., Errico, R. and Tai, K.-S. 2013. The influence of observation errors on analysis error and forecast skill investigated with an observing system simulation experiment. *J. Geophys. Res. Atmos.* **118**, 5332–5346. doi:10.1002/jgrd.50452
- Privé, N., Errico, R. and Tai, K.-S. 2014. The impact of increased frequency of rawinsonde observations on forecast skill investigated with an observing system simulation experiment. *Mon. Weather Rev.* **142**, 1823–1834. doi:10.1175/MWR-D-13-00237.1
- Privé, N. and Errico, R. M. 2013. The role of model and initial condition error in numerical weather forecasting investigated with an observing system simulation experiment. *Tellus-A* **68**. doi:10.3402/tellusa.v65i0.21740.
- Privé, N. C., Errico, R., Todling, R. and Akkraoui, A. E. 2021. Evaluation of adjoint-based observation impacts. *Q. J. R. Meteorol. Soc.* **147**, 121–138. doi: 10.1002/qj.3909
- Rienecker, M. and co-authors, 2008. The GEOS-5 data assimilation system - documentation of versions 5.0.1, 5.1.0, and 5.2.0. Technical report series on global modeling and data assimilation. 118 pp., NASA.
- Riishojgaard, L., Ma, Z., Masutani, M., Woollen, J., Emmitt, G. and co-authors. 2012. Observation system simulation experiments for a global wind observing sounder. *Geophys. Res. Lett.* **39**.
- Stoffelen, A., Marseille, G., Bouttier, F., Vasiljevic, D., de Haan, S. and co-authors. 2006. ADM-aeolus doppler wind lidar observing system simulation experiment. *Q. J. R. Meteorol. Soc.* **132**, 1927–1947. doi:10.1256/qj.05.83
- Timmermans, R., Lahoz, W., Attié, J.-L., Peuch, V.-H., Curier, R. and co-authors. 2015. Observing system simulation experiments for air quality. *Atmos. Environ.* **115**, 199–213. doi:10.1016/j.atmosenv.2015.05.032
- Vecchi, G. and Harrison, M. J. 2007. An observing system simulation experiment for the Indian Ocean. *J. Clim.* **20**, 3300–3319. doi:10.1175/JCLI4147.1
- Zhu, T., Weng, F., Masutani, M. and Woollen, J. 2012. Synthetic radiance simulation and evaluation for a joint observing system simulation experiment. *J. Geophys. Res.* **117**.