Technical
Memo

ECMWF
European Centre for Medium-Range
Weather Forecasts

916

Methods for assessing the impact of current and future components of the global observing system

Originally presented to ECMWF's Scientific Advisory Committee on 3 October 2022

Sean Healy, Niels Bormann, Alan Geer, Elias Hólm, Bruce Ingleby, Katie Lean, Katrin Lonitz and Cristina Lupu

April 2024

Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our web site under:

http://www.ecmwf.int/publications/

Contact: library@ecmwf.int

European Centre for Medium Range Weather Forecasts
Shinfield Park, Reading, Berkshire RG2 9AX, England

The information within this publication is given in good faith and considered to be true, but ECMWF accepts no liability for error, omission and for loss or damage arising from its use.

Abstract

We review the assessment of existing observations with Observing System Experiments (OSEs) and the Forecast Sensitivity to Observation Impact (FSOI) approach. Although care is needed when interpreting their results, the information they provide is largely consistent. The Ensemble of Data Assimilations (EDA) provides an affordable and manageable framework for simulating the impact of future observing systems. Recent experience comparing EDA predictions with the subsequent impact of real measurements gives us some confidence that, with appropriate interpretation and care, they provide useful information that can help guide the future evolution of the global observing system.

Plain Language Summary

This “Special Topic Paper” was originally presented to the ECMWF Science Advisory Committee (SAC) in October 2022. It has been reproduced here, with minor editorial changes, as an ECMWF Technical Memorandum to enable broader access to the document.

The central question addressed in the paper is: *How can we predict the potential impact of future observations on the quality weather forecasts produced with numerical weather prediction systems?* This is clearly fundamental when trying to plan how the global observing system (GOS) should evolve, but it is extremely difficult to address in practice. To provide appropriate context, this paper starts by reviewing how the impact of the current, real observations is assessed, emphasising that this apparently straightforward task requires considerable skill and care when interpreting the results. We then discuss the use of ensemble methods introduced by ECMWF in 2007 designed to *predict* the impact of the future observations on *theoretical* estimates of analysis and short-range forecast error *statistics*. The strengths and weaknesses of these ensemble methods are discussed, and examples using both current, real and future satellite observations are presented.

1. Background

This Special Topic Paper was originally presented to ECMWF's Scientific Advisory Committee (SAC) in October 2022. It has been reproduced here as an ECMWF Technical Memorandum, with minor editorial changes, to enable broader access to the SAC document.

The ECMWF Strategy 2021-2030 states:

> *"ECMWF will continue to strengthen its key role, in collaboration with space agencies and the WMO, to define and support long-term visions for global observing system developments."*

Observing System Experiments (OSEs) and Forecast Sensitivity to Observation Impact (FSOI) methods are used to assess the impact of current observations, and they can provide important insights into how the global observing system (GOS) should evolve (e.g. WMO, 2020). However, the long-term vision for the GOS also requires estimating the information content and potential forecast impact of future observing systems, sometimes before the instruments are developed. An assessment of this type will usually be based on simulated (or "synthetic") observations. The most established approach for estimating forecast impact with simulated observations is to run Observing System Simulation Experiments (OSSEs) (e.g. Errico and Privé, 2018). These are significant computational exercises, requiring both the integration of a high resolution "nature run" (NR), which is used as a proxy for the true atmospheric state, and the simulation of *all* existing observing systems.

Whilst ECMWF routinely supports OSSEs through the provision of the NRs, it does not conduct them. Running OSSEs at ECMWF would require significant resources. However, in recent years, we have developed an alternative, computationally cheaper method for assessing new observing systems, using an Ensemble of Data Assimilations (EDA) technique. This approach estimates how the new observations reduce the statistical uncertainty in the analyses and short-range forecasts.

Tan et al. (2007) pioneered the method, showing that the assimilation of simulated Aeolus measurements in an EDA system reduced the ensemble "spread", and that this spread reduction could be used to estimate the impact of the Aeolus measurements. The EDA approach was subsequently used to estimate how the impact of GNSS-RO scales with observation number (Harnisch et al., 2013), and this work informed the 20,000 occultation per day target, currently adopted by the International Radio Occultation Working Group (IROWG).

In 2021, the SAC noted:

> "*The SAC welcomes and supports the active role ECMWF plays in the preparation for future satellites, often in cooperation with satellite agencies, and evaluation of new data sets and their impacts. The SAC recognises the potential of using the EDA method to assess the contribution of existing and future observing system. However, like for other methods measuring impact, the interpretation needs to consider the limitations of the EDA calibration as well as the constraints of the configuration used for performing the evaluation of the existing or future observing system.*"

A key aim of this paper is to outline our current interpretation and understanding of the strengths and weaknesses of the EDA assessment technique, based on recent and ongoing studies. Clearly, it is important to engage with the space agencies on these matters when requested, but we must also be clear about the limitations in the information we provide.

Predicting the impact of a future observing system for NWP is very difficult with any method, because the overall signal will usually be relatively small when the measurements are added to the full observing system. In fact, it is recognised that assessing impact of ***real*** observations can now be challenging, because the NWP systems are already well constrained and skilful. Ideally, we want to demonstrate a statistically significant impact in the medium-range from a new observing system, but this can be difficult to achieve in a reasonable timeframe, so often a balanced judgment based on shorter-range forecast improvements and consistency with other observations, is required prior to operational use. It is useful to initially describe how real observations are evaluated, before explaining how we try to estimate the impact of a future system with the EDA.

In Section 2, we review how real observations are assessed, using both Observing System Experiments (OSEs) and the Forecast Sensitivity to Observation Impact (FSOI) approach. Both methods are routinely used to quantify the impact of components of the existing observing system, and they help support choices about the future evolution of the GOS based on current impact.

The main role of the EDA in the IFS is to provide flow dependent uncertainty information in the HRES system and contribute to the representation of initial condition uncertainties used in the ensemble forecasts. We review the key aspects of the EDA in Section 3, with a focus on understanding the sources of the EDA "spread", the current need to inflate the spread values when constructing the covariance matrices used in operations, and the possible reasons for this "under dispersion".

In Section 4, we present the use of the EDA for assessing new observing systems. The similarities with theoretical 1D-Var information content studies are noted. Previous EDA studies by Tan et al. (2007) and Harnisch et al. (2013) are briefly reviewed. Two recent studies funded by ESA, investigating the short-range forecast spread-skill relationship for real GNSS-RO observations (Section 4.3), and assessing the impact of a proposed constellation of microwave sounders on small satellites (Section 4.4), are then summarised. Importantly, both studies have compared the spread reductions achieved with simulated and real observations, and they have shown good consistency.

Section 5 is a summary.

2. Assessing the impact of existing observing systems

Assessing the impact of existing observations can provide important insights into how the global observing system should evolve. In this section, we review how we assess existing systems with using Observing System Experiments and the Forecast Sensitivity to Observation Impact method.

2.1 Observing System Experiments

An observing system experiment (OSE) measures the impact of a change in the real observing system. A **denial** experiment starts from the full observing system and removes one or more components. Denial experiments are used to assess the relative benefits of different types of observations. An **addition** experiment measures the impact of adding a new component to the existing observing system. These are used before deciding whether to activate a new observation type in the operational forecasting system.

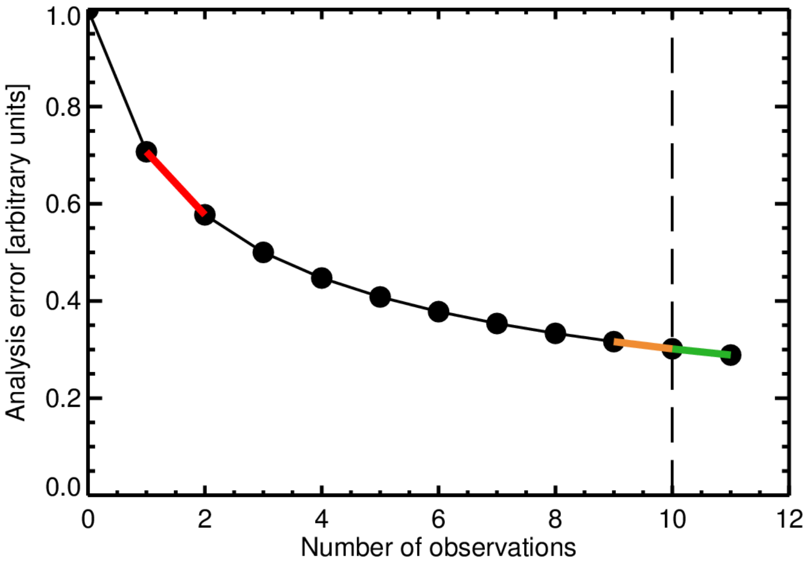

Figure 1: Standard deviation of Analysis error as a function of the number of observations. Here, the background error is 1 unit and each observation has an error of 1 unit, with no error correlations or biases. If 10 observations represent the current global observing system, then the orange line shows a denial experiment and the green line an addition experiment. The red line represents a low baseline addition experiment. As more observations are added, their individual impact reduces.

A fundamental difficulty with measuring changes in the observing system is that analysis and short-range forecast error statistics are expected to reduce following the inverse of the square root of the number of observations (See Figure 1, assuming all observations have the same error and are uncorrelated). A small modification within a much larger observing system typically makes only a small impact on the analysis and forecast error statistics, as illustrated in the addition and denial examples in Figure 1. This makes it hard to distinguish the signal of the change in the observing system from the natural chaotic noise in any weather forecast (Geer, 2016).

Some studies try to circumvent this by performing **low baseline,** poorer quality experiments, represented by the red line in Figure 1. This seeks to exclude most other observations in order to get a larger signal from adding the observing system of interest. A typical low baseline includes conventional data plus observations from one satellite, but incorrectly uses settings that have been tuned with the full observing system, such as the background error (e.g. Kelly et al., 2008). This can give insufficient weight to the observations under test, though Duncan et al. (2021) showed that for smaller, but still significant perturbations of the observing system (no microwave sounders), the issue of background errors is less important. The issue of an incorrect background error can also be overcome using an analysis reinitialization approach, where the background always comes from the full observing system (see Geer et al., 2014). However, these are not the only issues with the low baseline approach. A sub-optimal observing system, for example one with an uncorrected bias, might be able to show a beneficial impact on top of a low baseline system, but this impact is probably overestimated, and in the worst case the new data could degrade the analysis when added to a more complete observing system. Conversely, observations sensitive to nonlinear processes like cloud and precipitation may cause problems in a system where the background is too far from the truth, whereas they might be used more effectively when making small corrections in a more linear regime that has been found by using other components of the observing system. Other synergies between observations, such as the combination of limb (e.g. GNSS-RO) and nadir (e.g. radiosonde or sounder) measurements, may also be important. Assimilation experiments performed outside the main operational centres can also resemble a low baseline approach, since the experimentation may not include the full observing system and may not have access to a state-of-the art assimilation framework. For these and many other reasons, low baseline experimentation can be unrealistic, and it is most informative to measure the impact of an observing system in the context of all other observations in the best possible assimilation framework.

A potentially problematic approach used in some studies is to add observing systems progressively. This makes the impact of each observing system dependent on the order in which it was added to the system, as is obvious from Figure 1. The main context in which this is acceptable is in measuring the saturation of errors when progressively adding a set of interchangeable observing systems (e.g. Duncan et al., 2021).

Apart from the configuration of the experiments, there are many other challenges in interpreting OSEs. First is measuring the skill of the forecast or the quality of the analysis. It is hard to know how to trade off impacts at different forecast ranges, in different areas, or on different variables. Statistics can be aggregated to a single number (such as the Met Office NWP index) or a visual table (such as the ECMWF scorecard). However, it is important to understand the details, but these can amount to thousands of different plots. An alternative strategy is to concentrate on a single representative measure, such as the error in the 500 hPa geopotential height (Z500) in the medium-range. Outside the tropics the 500 hPa geopotential height measures the fundamental target of medium-range forecasting, which is the synoptic situation; verification of other fields such as temperatures and winds often give very similar results (Geer, 2016). Section 2.1.3 will discuss this further.

A second issue is finding a reference against which to measure the errors, given that the truth is not available. By definition the operational analysis using the full observing system is statistically the best estimate of the atmospheric state. However, using this to verify a forecast can be problematic due to the error correlations between analyses and forecasts. Verification 'against observations' sounds like a good thing but it means verifying against a partial subset of observations with random errors that are much larger than in the analysis, with more limited coverage, and with the systematic errors characteristic of that observation type. Third, the weather is a chaotic dynamical system in which small perturbations can grow nonlinearly into large errors in the subsequent forecasts. This makes it challenging to establish statistical significance, it means that experimental case studies (such as a single tropical cyclone) can be unreliable guides to true performance, and it complicates the link between the quality of the analysis and that of the subsequent forecasts, given that much of the error growth is driven from very localised areas.

An OSE is still the gold standard for understanding the impact of changes in the observing system, as long as it is made in the context of the full observing system and uses a state-of-the-art assimilation system. However, it is important to recognise the challenges that are described in more detail in the following subsections. (Some mathematical details on measuring analysis quality and forecast skill are given in the Appendix.)

2.1.1. Importance of the reference

Since there is no perfect reference, it is good practice to verify against both observations and analyses. In the research department it is common to use the statistics of background departures. Since these are calculated across the global observing system, it would be hard to find a more comprehensive set of observations for validation. The downside is that background departures are only routinely computed for the 12-hour forecast. It also means that verification is being done in the observation space, even if that is satellite radiances or bending angles, which need expertise to interpret. Direct observations of the forecast variables, such as radiosonde temperature or surface pressure, can be more easily used in observational verification at all forecast ranges, but these have the limitations of incomplete coverage (often dominated by highly populated areas of the northern hemisphere) and the relatively large errors in the observations themselves.

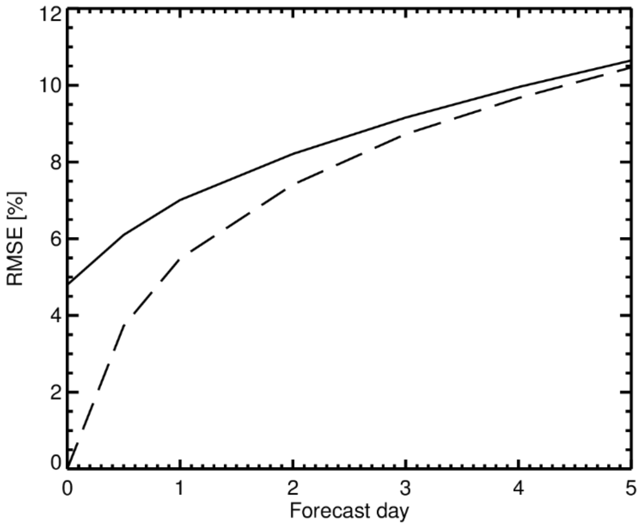

Figure 2: Forecast error in 850 hPa relative humidity in the tropics (20°N to 20°S) as a function of forecast day, computed against the operational HRES analysis (solid) or the experiment's own analysis (dashed). Against own-analysis, the apparent 0-hour forecast error is zero; 0-hour forecast error against the operational analysis was estimated from earlier studies to save time (Geer et al., 2010).

Verification against analysis has the advantage of complete coverage and of being based on the statistically best estimate of the atmospheric state. However, the results can be dominated by the error covariance term between the errors in the forecast and the errors in the reference (see also appendix for further details). Hence the choice of analysis to use as a reference is important. An extreme example of the problem is seen in the humidity field in the tropics (Figure 2). At a 12-hour forecast range, the RMSE is around 40 % smaller when computed against the experiment's own analysis than when it is calculated against the operational analysis. Since the 12-hour forecast is used as the background in the data assimilation, the resulting analysis is strongly correlated with the 12-hour forecast. Another way to think about this is that 12-hour RMSE computed against own analysis is also a measure of the size of the data assimilation increments. The effect of adding new observations is often to increase the size of the increments, reduce correlations between forecast and analysis, and hence to apparently increase the size of the RMSE (e.g. Bouttier and Kelly, 2001; Geer and Bauer, 2010).

The operational analysis used in the verification in Figure 2 is made using the operational 12-hour forecast. This has a different realisation of the forecast error, and hence is less correlated with the 12-hour forecast in the experiment. The 0-hour forecast validation against the operational HRES analysis (strictly, analysis vs. analysis verification) is similar to comparing two different members of the EDA ensemble, so the apparent error does not reduce to zero. The problem of using the operational analysis as the reference is that it can favour experiments that use the same observing system as the operational system. Systematic changes in the analyses and forecasts introduced by observations could be particularly problematic: for example it is thought that aircraft observations can bias upper-tropospheric temperatures in the analyses and short-range forecast, creating a systematic error with a spatial pattern that follows the main commercial airways over the US, Atlantic and Europe. Through patterns of systematic error, and possibly simply because similar observing systems make corrections to the analyses and forecasts in similar ways, verification against operational analysis can artificially reduce the RMSE in favour of similar observing systems.

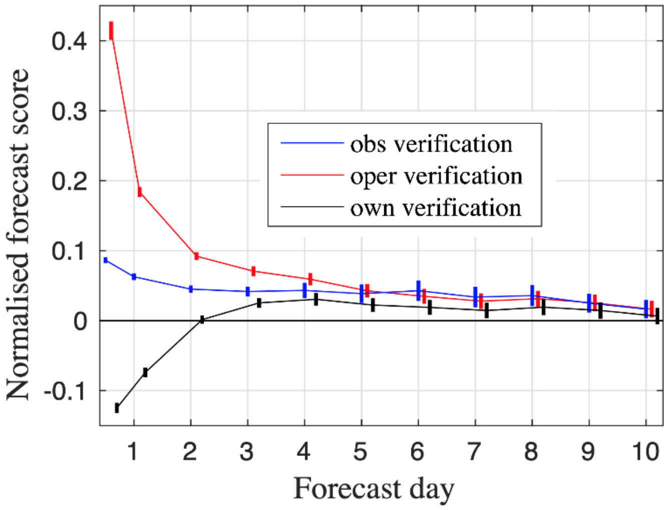

Figure 3: Normalised change in the RMSE vector wind forecast error at 200 hPa in the northern extratropics resulting from a denial of all conventional observations. For three different verification references, the operational analysis, radiosonde observations, and own analysis, the results are completely different in the short-range. Error bars represent the 95 % confidence interval following Geer (2016). Reproduced from the right hand panel of Fig. 3, Bormann et al., 2019.

For midlatitude dynamical verification, the issue is not as extreme as in the tropical humidity, but it is still significant. In the context of observing system denials, Lawrence et al. (2019) and Bormann et al. (2019) examined changes in RMSE with three possible references, own-analysis, operational analysis, and observations (e.g. Figure 3). For the 12-hour forecast in this example, the apparent increase in RMSE is 40 % when using the operational analyses as a reference, and just 10% against observations. Even at forecast day-2, verification against operational analysis gives twice the forecast degradation as measured against observations (10% vs. 5%). Though the use of the observational reference has its issues, this suggests that verification against the operational analysis gives an inflated estimate of the degradation caused by removing observing system components, out to at least day-3. As the forecast range increases, due to rapid error growth in the midlatitudes, the true errors in the forecast become relatively large compared to the errors in the reference or the error correlation term (Eq. A3, see Appendix) so the choice of reference becomes less important in the longer-range forecast. For the most reliable forecast evaluation at short forecast ranges, it is typical to favour observation-based measures as the most reliable verification reference.

2.1.2. Growth of forecast errors

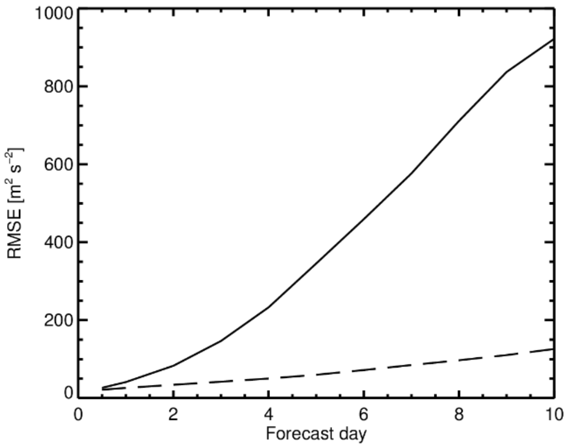

Figure 4: RMSE in geopotential forecasts at 500 hPa in the northern midlatitudes (solid) and in the tropics (dashed).

The atmosphere is a dynamic system with sensitive dependence on initial conditions, in other words a chaotic system. In the midlatitudes, baroclinic errors grow rapidly over the first 10 days of the forecast (Figure 4, solid line) and start to saturate beyond this. However, in the tropics, error growth is much slower but shows little sign of saturation (Figure 4, dashed line). As discussed, this means that true forecast errors in the tropics do not always dominate the errors in the reference or the covariance term between the forecast error and the reference error (Eq. A3, see Appendix). Also, the bias term can be much more important (Eq. A2). A consequence of chaotic error growth, particularly in the extra-tropics, is that even the smallest perturbation in the system can grow into a large synoptic-scale error. The EDA uses this effect to generate multiple different realisations of the possible analysis using perturbed observations and a perturbed model, and even a purely technical perturbation to the system can generate a similar effect (e.g. Geer, 2016).

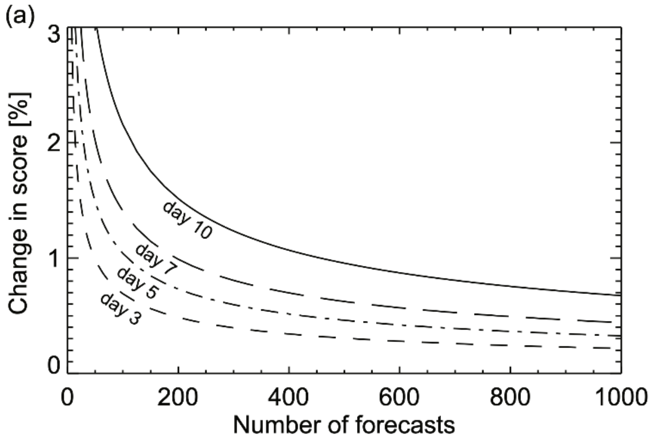

Figure 5: Size of the 95% statistical significance range for 500hPa geopotential height forecast in the extra-tropics, as a function of the number of forecasts in the sample, and of the forecast range. The significance range is expressed as a percentage change in the RMSE, normalised by the RMSE in the control experiment. The impact of adding one AMSU-A to the full observing system is around a 0.5 % reduction in the RMSE at day-5, meaning that at least 500 separate forecasts would be needed to measure this change with statistical significance. Reproduced from figure 15a of Geer (2016).

One of the difficulties of forecast verification is deciding whether a difference in the forecast quality has come from the change being tested, such as addition of a new observing system, or whether it is just the consequence of chaotic error growth. This is the role of statistical significance testing. However, chaotic errors are often large compared to the small signals being measured in an addition or denial experiment. This chaotic noise can be reduced by running experiments for long periods; Figure 5 shows how the impact of a single AMSU-A, which reduces forecast error by around 0.5% in the medium-range, would need to be aggregated from around 500 separate forecasts to attain statistical significance at forecast day-5 (Geer, 2016).

A final consequence of chaotic error growth is the importance of small areas of high sensitivity, often mapping onto baroclinic features. The existence of these sensitive areas is clearly shown by adjoint modelling (e.g. Rabier et al., 1996). However, the existence of sensitive areas may make short-range verification an unreliable predictor of changes in the medium-range forecast skill. Imagine a new observing system that predominantly improves the analysis in the subtropical subsidence regions. These are areas which seem to have very small influence on the subsequent forecast; the short-range verification might be improved, but if the new observing system was unable to improve the analysis in areas of rapidly growing forecast errors, it might not have any effect on the medium-range forecasts. A similar effect is seen in the decorrelation of measured forecast skill at longer ranges: just because the day-3 forecast is good does not mean the day-10 forecast will also be good (e.g. Geer, 2016). This is mainly a caution against placing too much weight on the short-range verification, which is often a temptation if statistical significance cannot be established in the medium-range.

2.1.3. Summary

For decades, medium-range weather forecasting has used the midlatitude geopotential height errors at 500 hPa as a main indicator of synoptic forecast quality. This can seem overly reductive. However, these errors are easy to measure, since they far outweigh any errors in the reference, and they rapidly outgrow the error covariance term that makes the choice of reference so important in the earlier forecast range. Finally, the medium-range forecast is an effective filter, taking whatever information is useful from the observations (in areas where the errors are rapidly growing) but ignoring observational information in areas that don't matter (in areas of decaying errors). It is hard to measurably improve the medium-range synoptic forecast, due to the size of the chaotic variability, and due to the relatively small signals involved in high-baseline addition or denial experiments (Figure 1). But for these reasons, a statistically significant reduction in the day-5 extratropical 500hPa geopotential height RMSE is an unambiguous signal that an observing system has a beneficial impact. Verification of the short-range forecasts, and of areas or fields with more complex error patterns, such as the stratosphere, the tropics, or clouds and humidity generally, is also important, but needs to be approached with much greater caution.

2.2 Forecast Sensitivity to Observation Impact (FSOI)

Forecast Sensitivity to Observation Impact (FSOI, Langland and Baker, 2004) is an adjoint or ensemble-based method that enables the simultaneous assessment of how different groups of observations contributed to the reduction of a globally integrated, quadratic function of short-range forecast errors. It is argued that the reduced error in a 24-h forecast from the analysis, compared to the 36-h error forecast from the background, is due to the work done by the observations to improve the description of the atmospheric state. Using the adjoint (in the ECMWF implementation) of the forecast model, the forecast error differences, usually expressed as a total energy norm, are traced back in time (linearly) and attributed to individual observations according to their weight in the analysis, and relevance to the chosen metric. Given that FSOI relies on the use of a linearised version of the forecast model, it is only valid to evaluate short-range forecasts (0 to 48 hours, Janisková and Lopez, 2013). As discussed in the previous section, FSOI results at short-range can only be considered a partial indicator of medium-range forecast impact.

The choice of forecast error measure, including the verifying reference, is an important aspect of the FSOI application, and it needs to be decided at the start of the FSOI calculations. Most NWP centres use a dry or moist total energy norm as an overall measure of forecast skill, obtained by verifying against the own analysis. The choice of dry or moist energy norm can have some impact on results and conclusions subsequently reached (Marquet et al., 2020), as it leads to different weighting of forecast errors across variables and levels. The choice of verifying reference can also have a significant impact on the results, as errors in the verifying analysis can confuse the interpretation of the impact of the observations. This can be due to biases in analysis and forecast, as well as correlated errors between analysis and forecast, and this problem is very similar to the issues highlighted earlier in the context of short-range forecast verification of OSEs. While the use of own-analysis verification is most common for FSOI applications, Todling (2013) and Cardinali (2018) tested instead the use of an observation-based forecast error metric, to avoid some of these issues. Necessarily, results differed from the analysis-based FSOI, not least due to the geographical and geophysical sampling of the observing system and giving more weight to the stratosphere because of the satellite information the. At ECMWF, the default configurations used in operations applies the dry total energy norm to assess the impact on T+24 forecasts (Cardinali, 2009), using own-analysis verification, and the system is run twice daily based, using the long-window configuration of the HRES system.

Since FSOI provides an impact measure for individual observations, it can be aggregated in any way, and this is a particularly attractive feature. It allows affordable detailed analyses for multiple observation types, and also subsets of observations, such as individual stations or channels of an instrument, specific geographical regions, observed variables, etc. Results are typically reported either as total FSOI for a particular observation set, or as FSOI per observation. Some aggregation is needed, because individual values are noisy. In this context, FSOI suggests that just over 50% of the observations have positive impact, reflecting the statistical nature of FSOI (Lorenc and Marriott, 2014). As a result, a thorough interpretation of FSOI results should include statistical significance testing.

FSOI can be particularly useful for assessing the value of remote observations which are considered meteorologically important, but are difficult to assess in OSEs, as the size of the impact signal will be comparatively small and localised. For example, in support of the WMO Systematic Observations Financing Facility (SOFF https://public.wmo.int/en/our-mandate/how-we-do-it/development-partnerships/Innovating-finance), we showed that the Synop FSOI per datum in the Pacific Islands is more than 4 times the impact per datum achieved in the US or Europe (Figure 6), and this is useful information to justify continued investment in such observations. Of course, high density observations over Europe have value for short-range European forecasts, and more observations just upstream, over the Eastern North Atlantic, would be very useful.

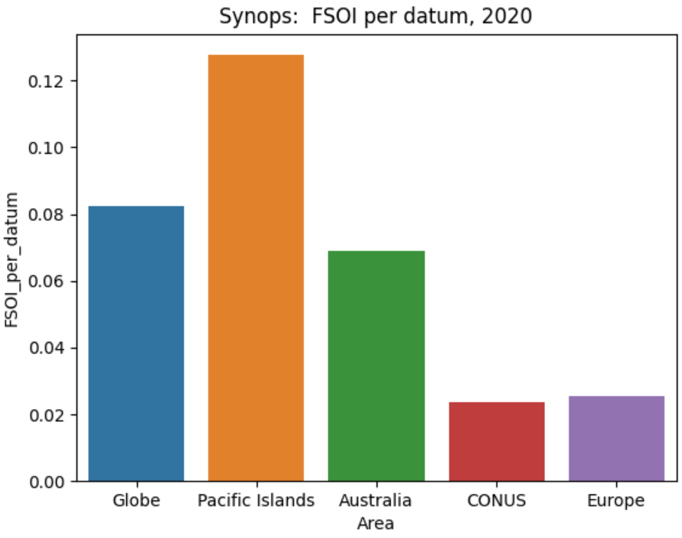

Figure 6: FSOI per datum (J/kg, positive values denote better forecasts) for SYNOP stations in 2020 - various regions (CONUS denotes the Contiguous USA).

FSOI also enables us to monitor and summarise how the contributions of different observing systems evolve over time. For example, Figure 7 shows such a plot for 2020 to June 2022. There are several features of interest: a) the increase of GNSS-RO impact (brown) with the implementation of COSMIC-2 assimilation in March 2020 and Spire data in May - removed at the end of September (the Spire data were made temporarily available because of the Covid pandemic), b) the drop in aircraft numbers and impact (cyan) in March 2020 followed by a slow recovery (see Ingleby et al., 2021), c) the impact of Aeolus wind data (yellow dashed line, Rennie et al., 2021), which was about 4% of the total when introduced in January 2020, which is considered very large for a prototype satellite, reflecting the relative sparsity of wind data, especially in the tropics. In Figure 7 the 20-day mean values still show some noise, partly due to fluctuations in observation numbers, but much of it can be considered as random. Overall, in situ observations (densest in the northern hemisphere) tend to have slightly larger FSOI in the boreal winter than the summer, reflecting slightly larger background errors in winter. Over the last eight or so years (not shown) there has been a strong increase in the impact of microwave data, especially the humidity channels (MWWV, black in Figure 7) with the addition of several instruments and the move to all-sky radiance assimilation (Geer et al., 2017).

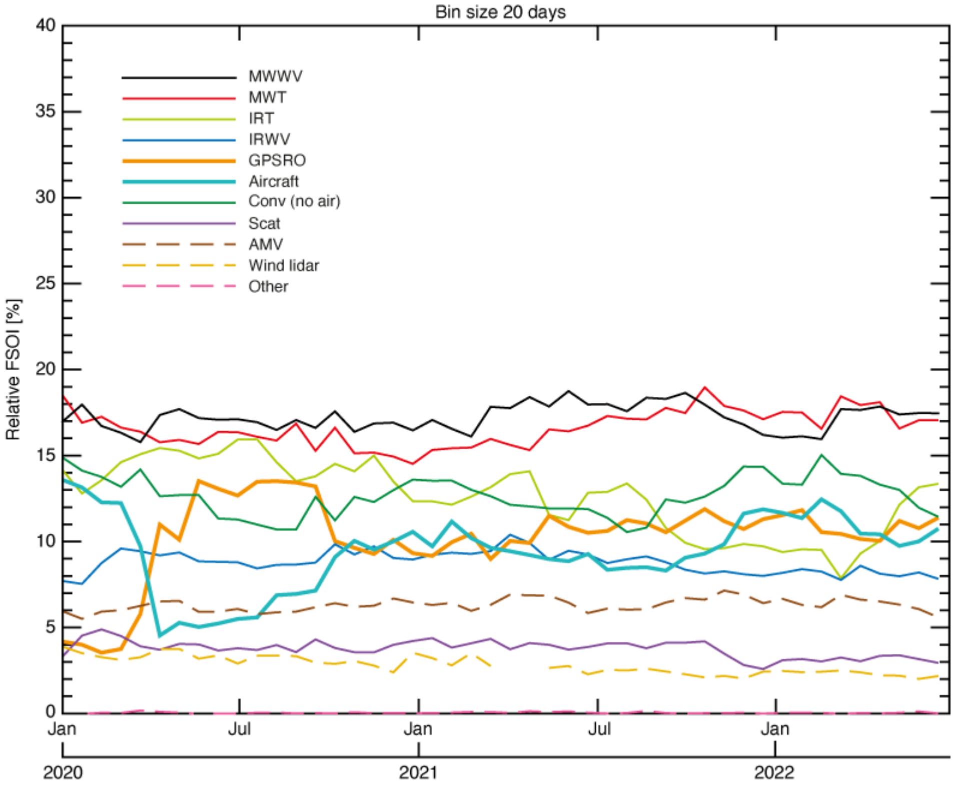

Figure 7: Relative impact of different observation subsets (% of FSOI), 20 day mean.

Conceptually, FSOI and OSEs are very different, and they are usually seen as complementary pieces of information. OSEs measure the cumulative impact of removing (or adding) observation subsets, including any effect via adjustments to bias correction, whereas FSOI gives results linearised about the 'all observations' states. Therefore, FSOI does not measure the impact of removing the observation type from the system. For example, if an observing system provides 20% in FSOI, this does not mean that forecast errors will increase by 20% if it is removed. The DA system will automatically re-tune, giving more weight to other observations by taking account of the degradation of the background, and the loss of skill may be smaller. The different behaviours of OSEs and FSOI have been highlighted by Eyre (2021). The OSE results are more resilient to the removal of observational information from well-observed variables, but not when denied from poorly observed variables.

As in the case of OSEs, FSOI results will depend on the maturity of the NWP system used, including the number and diversity of observing systems assimilated. Two different NWP systems may give very different results. In the case of FSOI, an important caveat is that results are very sensitive to assigned observation errors.

Overweighting an observation type can lead to a misleadingly large FSOI for that observation type. In situations where an observation type is over-weighted to the extent that it gives negative impact in an OSE, FSOI may misleadingly suggest particularly large positive influence from this observation type relative to others (Lupu et al., 2015). FSOI therefore requires careful analysis to avoid misinterpretation, and ideally it should be used alongside OSEs.

In general, FSOI has a less important role at ECMWF when assessing whether a new observing system should become operational, and the relevant OSEs are used instead for this decision. FSOI is used primarily for diagnostic purposes. For these, either the operational database of FSOI values for each observation is available, or dedicated RD experiments can be run. FSOI is not calculated by default for RD experiments.

3. Ensemble of Data Assimilations

We use the EDA system in the assessment of new observations, but the EDA is primarily designed to provide flow dependent uncertainty information. This section outlines the EDA method and how it is used in the main applications.

3.1 Method

The Ensemble of Data Assimilations (EDA) (Isaksen et al., 2010; Bonavita et al., 2016) has the dual role of providing flow dependent covariance estimates for the hybrid 4D-Var system, and it contributes to the estimation of the initial conditions used in the ensemble forecasting system (ENS). In the EDA, each perturbed ensemble member performs an independent 4D-Var analysis update (Courtier et al., 1994; Rabier et al., 2000). The EDA is a variational implementation of a Monte Carlo approximation to the standard Kalman filter (Houtekamer and Mitchell, 1998), also known as the perturbed observations ensemble Kalman filter (EnKF), where observations in each member are perturbed according to their assumed observation error covariances.

3.2 EDA Spread: Source and reliability

In theory, if all sources of uncertainty in the assimilation system are accurately sampled, the analysis and background covariance matrices, given as (**A**, **B**), respectively, sampled from the ensemble members are, on average, unbiased estimators of the error covariances of the system. In practice, the “raw spread” calculated directly from EDA members is lower than the corresponding error statistics estimated from cross validation or spread-skill relationships (see Bonavita et al., 2016). This “under dispersion” of the EDA is common to most/all operational ensemble data assimilation systems, and most employ methods for covariance inflation, typically multiplicative or additive inflation (see review by Houtekamer and Zhang, 2016). At ECMWF, however, there is no inflation of the cycling EDA members, and the focus is on improving the modelling of the sources of uncertainty.

The “reliability” characteristics of the ECMWF EDA are thus dependent on our ability to model the sources of uncertainty, including observation errors, model uncertainty, boundary and forcing term uncertainty and system uncertainty (see Houtekamer and Zhang, 2016, Table 4, for a list of error sources). Observation uncertainty is modelled by the perturbed observation approach that adds random perturbations with the standard deviation of the assumed observation error to the observations used in each member. The model uncertainty parametrization is identical to the one used in the ensemble forecasts (Lock et al., 2019), and it is added to the model of each ensemble member. Boundary and forcing terms are only partially represented with climatological SST perturbations. Land, sea ice and snow-covered surfaces are not explicitly perturbed. System uncertainty is not explicitly accounted for, but we attempt to reduce its amplitude by applying the same 4D-Var system in each member as in the unperturbed analysis, apart from the differences in resolution and number of outer loops for cost considerations.

There are several sources of bias in the current uncertainty estimates. Observation error amplitude and correlations are sometimes simplified. For example, we omit the spatial correlation of AMV observation errors for computational cost reasons and increase their standard deviations to compensate. More generally, it is conceivable that the lack of explicit spatial observation error correlations in other dense observing systems (e.g., microwave and infrared sounders) could be a significant source of missing variance in the EDA error budget (Liu and Rabier, 2002). Model error uncertainty is calibrated for performance in the ensemble forecast system, but in some areas, terms are omitted or reduced. For example, we reduce the perturbations in the boundary layer for stability, and there is also a lack of explicit perturbations to the surface fluxes (Leutbecher et al., 2017). Further omissions are a lack of perturbations in the land surface models, which will reduce boundary layer spread in the atmospheric boundary layer.

The under dispersion of the cycling EDA system means we need to apply inflation to the covariances **A** and **B** calculated from the raw EDA at the point where they are used. For the background errors, **B,** currently a global inflation factor of 1.34 is applied when calculating background errors standard deviations from raw EDA spread values. This factor is heuristically derived from testing which covariance inflation gives the best 4D-Var performance, both in terms of observation departures and medium-range forecast scores, and it changes over time with updates to the system. This multiplication is applied inside the analysis, where there is also a further modification of the amplitude to account for the difference in global variance at different inner loop truncations.

A remaining question is whether the relative spread changes we see by comparing EDA's run in a simpler configuration, for example the lower resolution and fewer members used when assessing new observations, are representative for changes we would see in the full resolution system. This has been addressed in the context of designing lower resolution and fewer members EDA experiments, to test the impact of proposed changes to the full resolution system (Lang et al., 2019). In the majority of cases, the changes in the low-resolution system are representative, with only a minority of changes requiring full resolution testing. For EDA observation impact studies, we can likewise use fewer members. Lower resolution will likely give similar results for most observation types, with the provision that certain observation types are more sensitive to resolution and need more resolution, for example TCo639 (18 km spacing) rather than TCo399 (29 km spacing).

There are currently several areas of work being considered to improve the reliability of the EDA. These include improved model uncertainty parameterizations, better modelling of observation errors including horizontal correlations, and a reduction of the resolution gap between the EDA and the high-resolution analysis. By addressing the sources of unreliability while not inflating the cycling EDA, we aim to stepwise improve the reliability of the EDA.

3.3 Use of EDA in Ensemble Forecasts

Buizza et al. (2008) examined ensemble initial conditions based on the EDA. They found that initial perturbations based solely on the EDA did not generate sufficient ensemble spread to obtain reliable ensemble forecasts. Their study also looked at the combination of initial perturbations based on the EDA and perturbations based on the leading initial singular vectors (SV), which led to a larger ensemble spread that better matched the growth of the error of the ensemble mean. The combination of EDA and singular vector perturbations has been used operationally since 2010, and the amplitudes of the singular vector perturbations have not changed significantly since 2011. The need for this kind of inflation may disappear with future improvements in the EDA reliability, but it currently has implications for using the EDA to estimate the medium-range impact of new observations (Section 4.5).

4. Using the EDA to estimate the impact of new observing systems

As discussed in Section 3, the cycling EDA system is designed to provide flow dependent covariance information. Currently, like most operational ensemble-based DA systems, it is under dispersive, so in practice the “raw” covariance matrices have to be inflated for operational use in the 4D-Var, and singular vectors are still required in combination with the EDA when initialising the ensemble forecasts.

In this section, we outline how the EDA is used to assess the impact of new observing systems and provide examples from previous and ongoing work. In this application, we are investigating how the new simulated observations change the estimates of the analysis and short-range forecast error *statistics* provided by the EDA These changes in the error statistics are usually averaged over both large spatial scales and over a period of around a month, so it is a large scale, statistical measure of the observation impact. The EDA method does not provide a useful framework for assessing how the new observations improve forecasts of specific, high impact weather events.

Ideally, we would like to estimate the impact of the observations on the medium-range forecasts, but the EDA observation studies usually focus on the analysis and short-range, because the observation impact is clearest there. Currently, the EDA method is not ideal for the assessment of medium-range impact. It can be extended to longer forecast ranges (Section 4.5), but the interpretation of these results is more complicated, because the singular vectors needed in the ensemble system to achieve reliability are not included in the EDA.

4.1 Interpretation, link to OSSEs and simulation

4.1.1. Information content interpretation

The use of the EDA for assessing new observing systems can be interpreted as a 4D-Var information content approach, analogous to 1D-Var studies (e.g. Rodgers, 2000). It provides a *theoretical* estimate of how the (cycling) analysis and short-range forecast error statistics (**A** and **B**, respectively) should respond to assimilating a new set of observations. In the linear limit, these changes in the error statistics are dependent on the assumed observation error covariance, **R**, rather than the values of the new observations, **y.**

Clearly, the EDA must provide a *theoretical* estimate of the uncertainty reduction because the simulated measurements can reduce the spread values without improving the estimate of the atmospheric state. However, it will be demonstrated that real and simulated COSMIC-2 measurements produce very similar spread reductions in the EDA (Section 4.3.1). The key requirements are producing simulated observations with realistic short-range forecast departure statistics (o-b), and using an accurate estimate of the observation error covariance matrix, **R** (Linearity assumptions can break down if an inappropriate **R** matrix is used in the EDA).

As the EDA is a cycling system, it estimates the accumulated impact from both the previous and the latest observations. The cycling also means we can estimate the impact of mass measurements on the wind field, for example how the GNSS-RO or microwave sounder measurements impact the tropical wind uncertainty estimates. This is not possible in a 1D-Var study.

4.1.2. Relationship to OSSEs

There are obviously many computational and technical differences with OSSEs, and we do not discuss OSSEs in detail here (see, Errico and Privé, 2018). However, a key difference is that OSSEs can provide direct estimates of analysis and forecast error vectors, $\boldsymbol{\varepsilon}_{\mathbf{a}}$and $\boldsymbol{\varepsilon}_{\mathbf{f}}$, respectively. This is because the true state is defined by the nature run. In OSSEs, covariance matrices, like **A** and **B**, can then be computed from the statistics of these error vectors. In contrast, the EDA is a cycling, error propagation model that attempts to estimate the **A** and **B** matrices directly, without reference to a true state.

OSSEs are computationally expensive when compared to the EDA, but it can be argued that they should potentially provide a more comprehensive assessment of the observation impact. We would argue that having access to both EDA and OSSEs when assessing a new observation should be beneficial, because the impact will differ from system to system. On the interpretation of OSSEs, Errico and Privé (2018) make many of the same points made in Section 2 for OSEs. In particular, they note that OSSEs provide statistical information on the observation impact. The impact of the new observations when added to the full observing system may be "modest", but using a reduced observing system will provide misleading results. Case studies from OSSEs do not generally provide robust statistics from which to draw confident conclusions. "Legitimate" OSSEs are unlikely to provide "extraordinary impacts", and decision makers should have reasonable expectations about the new data.

4.1.3. Method: simulation of observations used in EDA

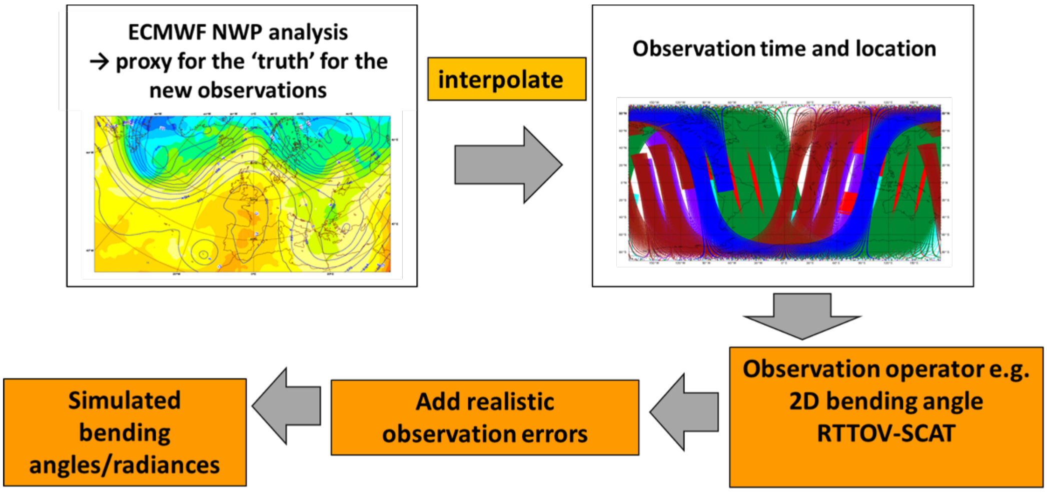
ECMWF NWP analysis
→ proxy for the 'truth' for the new observations
interpolate
Observation time and location
Observation operator e.g. 2D bending angle RTTOV-SCAT
Add realistic observation errors
Simulated bending angles/radiances

Figure 8: Typical simulation system used to produce observations for the EDA computations. Note that errors are added both in this simulation step and later during the EDA.

The EDA computations require a simulated dataset generated with a system illustrated in Figure 8. There are three main steps. High quality atmospheric state information is interpolated to the observation locations. A forward model maps the state information to observation space. Noise is added to the simulated data and it is stored ready for assimilation alongside real data.

There have been differences over time in how this simulation system has been implemented at ECMWF as understanding has evolved. For example, Tan et al. (2007) used Met Office short-range forecast information to ensure that the (o-b) departures were realistic. Subsequent studies used operational ECMWF analyses, but at higher resolution than the EDA. Harnisch et al. (2013) used a 2D bending angle operator to simulate the data, but a 1D operator to assimilate it, in order to introduce a realistic forward model error. The recent microwave sounder study in Section 4.4 uses the same observation operator (RTTOV-SCATT) to simulate and assimilate the radiances, but with different surface emissivity information over land and sea-ice. The advantages of using different forward models to simulate and assimilate the observations are not clear.

There are still some open questions on the importance and role of the noise added in the simulation step. We have to consider whether it should include just instrument noise, **E**, or both instrument noise and representation error, **R=E+F**, and how to model these. The GNSS-RO work usually adds perturbations that explicitly include both components, whereas for the MW sounder work it was found that realistic background departures can be achieved for the simulated observations when only instrument noise is explicitly considered, with representation error arising implicitly in the observation simulations. Introducing realistic biases for observing systems that are likely to require bias correction also needs some consideration.

4.2 Some previous EDA studies

4.2.1. Early Aeolus studies

The application of EDA techniques for assessing new observing systems was introduced by Tan et al. (2007) to estimate the potential impact of Aeolus line of sight (LOS) wind profile information in the ECMWF system. Their key insight was that the EDA computations – unlike OSSEs – could be performed with a mixture of both real and simulated measurements. This means that only the new observing system needed to be simulated.

It was recognised that OSSEs can potentially provide an estimate of the "absolute observation impact", because the simulated truth is known from the nature run (NR). In contrast, they noted that estimating the absolute impact of an observing system from the EDA computations would require careful calibration. However, it was argued that when comparing two observing systems, absolute calibration is unimportant, as it simplifies to a common rescaling of the spread values.

The study mainly focussed on the impact on analyses and short-range forecasts, but showed a zonal wind spread at day-5. They emphasised potentially large impact in the tropics. The possible reasons for the EDA under dispersion were discussed, where it was noted the model error was not explicitly accounted for in the EDA at the time, and that spatial correlations of observation errors were only included when assimilating SATOB AMV wind information.

It is difficult to compare EDA results from 2007 (investigating a period in 2003) with recent OSEs directly, but Rennie et al. (2021) have subsequently demonstrated significant impacts assimilating Aeolus measurements in the ECMWF system. The latest extended experiments, now spanning over 2 years using reprocessed Aeolus data, show statistically significant improvements in the northern/southern hemisphere Z500 out to day 4, and improved tropical winds out to day 8-9 (e.g., 200, 100 hPa).

4.2.2. GNSS-RO impact with observation number

Harnisch et al. (2013) applied the EDA method to estimate the impact of increasing the number of GNSS-RO measurements. The question at that time was whether there would be additional benefit from increasing the measurement numbers above the present levels of ~2500 occultations per day. Poli et al. (2008) had already demonstrated that the GNSS-RO impact on humidity and geopotential forecast errors was reduced when only half of the available GNSS-RO measurements were assimilated, but it was not clear how the impact might scale beyond the current numbers, or where "saturation" of impact might occur.

The EDA experiments covered the period July-August, 2008, and were performed for the ~2500 real GNSS-RO observations per day available at that time, and 2000, 4000, 8000, 16000, 32000, 64000 and 128000 simulated GNSS-RO measurements per day.

The large range of the GNSS-RO numbers was intended to investigate the possible onset of saturation of GNSS-RO impact, but the key result from the study was that the EDA spread values continued to reduce, even when moving from 64,000 to 128,000 simulated observations (Figure 9). (Note that the economic case for extending the GNSS-RO numbers to 128,000, versus other possible non GNSS-RO changes to GOS, was not considered or discussed in this work.)

The failure to achieve saturation was explained by the assumed GNSS-RO uncertainty model, where all vertical and horizontal GNSS-RO error correlations were assumed to be zero. It was argued that even perfectly correlated observations, with uncorrelated observation errors, will appear to provide useful information in the EDA system, because the additional observations effectively reduce the random measurement errors through repetition. It was also recognised that the concept of "saturation" was ambiguous, particularly when assimilating observations with uncorrelated errors. For example, in the idealised scalar case shown in Figure 1, the standard deviation of the analysis error scales as $1/\sqrt{1+n}$, and this will continue to fall as more data is added, albeit with a reduced rate, so "saturation" is difficult to define objectively.

The study concluded that there was a good case to increase the GNSS-RO numbers, but also encouraged work on OSSEs to test this further. Recent GNSS-RO OSSEs by Privé et al. (2022) have now shown that the GNSS-RO impact does not saturate at 100,000 occultations per day in their system. The Harnisch results informed an International Radio Occultation Working Group (IROWG) target of 20,000 occultations per day. Recent results showing the benefits of assimilating additional GNSS-RO from both COSMIC-2 and Spire suggest it was a reasonable target (Figure 7, Figure 10).

However, it must be acknowledged that if we reran the Harnisch study today, we would not achieve the percentage reductions in spread shown in Figure 9, which are only valid for the NWP system and observation usage in July-August, 2008. This is because the background EDA spread values have reduced since 2008, primarily because the total number of observations assimilated has increased over the period (although other improvements to the NWP system will also have an impact). In 2008, the number of observations assimilated per day was below 10 million, but by 2020 it exceeded 30 million. The "no RO" 12-hour temperature spread values at 100 hPa are now around 0.15-0.2 K lower than in the 2008, which is a ~20-30 % spread reduction. We have also increased the assumed GNSS-RO error statistics used to assimilate the data. Consequently, the additional GNSS-RO measurements will have a lower information content today, in both an absolute and relative sense, because other observations are providing more information and GNSS-RO slightly less. This highlights potential problems that will arise if we try to compare spread reductions from different EDA experiments. The information content of the new observations will depend on the quality of the NWP system, resolution and observations used in the control experiment, and this must be common in any comparisons.

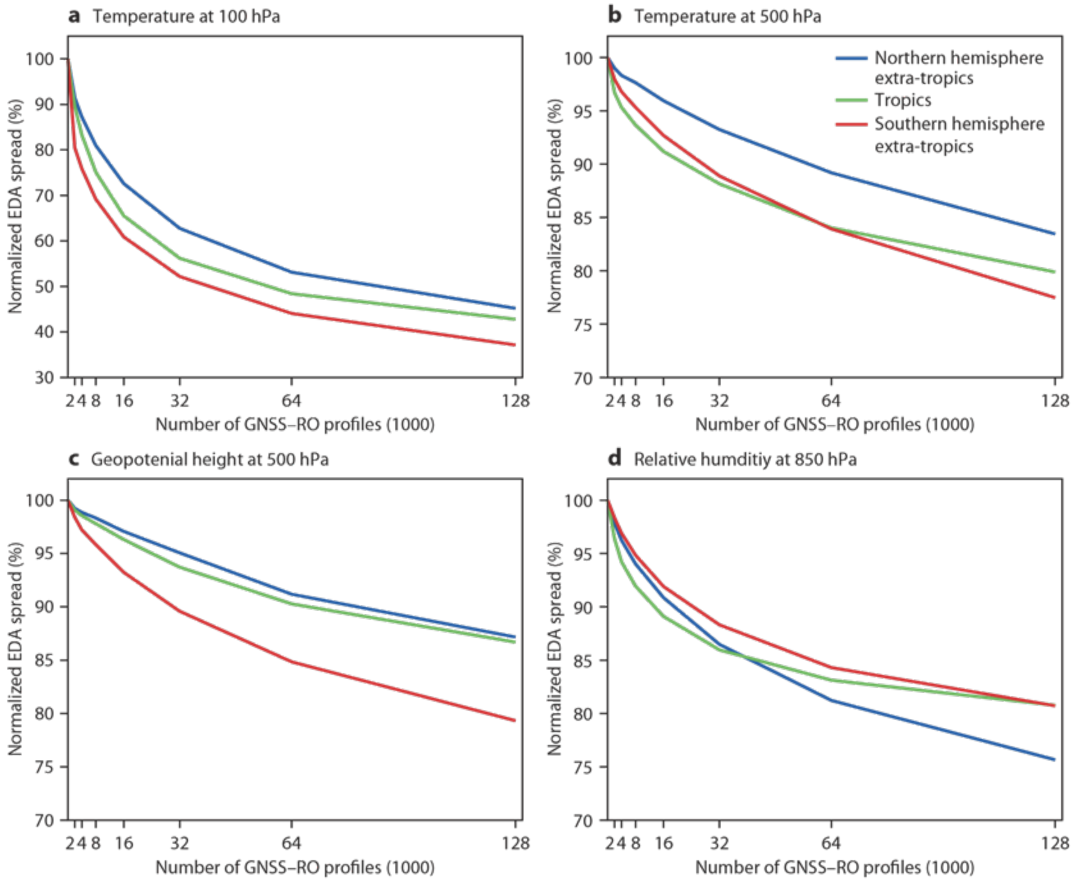

Figure 9: The change in normalised (divided by the "no RO" spread values) analysis EDA spread as a function of the number of simulated GNSS-RO measurements, from Harnisch et al. (2013).

4.3 Recent application to GNSS-RO

We have recently completed an ESA study in collaboration with the Met Office and EUMETSAT, testing the combined impact of (~) 3900 COSMIC-2 and 6800 Spire GNSS-RO measurements per day in two NWP systems (Lonitz et al., 2021). The total number of GNSS-RO was ~13,000 profiles per day. The experiments covered the period January 1 to March 31, 2020. The main aim of the study was to determine the *actual* forecast impact of this *real* data, to help inform agency commercial data buy decisions. It was also used as an opportunity to compare the impact of simulated and real measurements in the current EDA system, and to examine the real EDA spread-skill relationship, as the GNSS-RO numbers increased significantly.

Figure **10** shows the combined impact of the COSMIC-2 and Spire data on the Z500 RMSE scores in the northern and southern hemisphere extra-tropics, verified against both own analysis (a) and operations (b). The impact of GNSS-RO has roughly doubled at the short-range with the addition of the new data. This is comparable to spread reductions given in the Harnish study, which showed a factor of ~3 enhancement in the SH Z500 impact at short-range, when moving from 2000 to 16000 simulated observations per day (See their Figure 6b). However, as discussed in Section 2.1.1 (See Figure 3), the magnitude of the GNSS-RO impact at short-range is sensitive to the choice of verifying analyses, particularly the degradation seen when removing all GNSS-RO from the system, because these observations are used in the operational analyses. This complicates direct comparisons between EDA spread reductions and the improved forecast scores computed in OSEs.

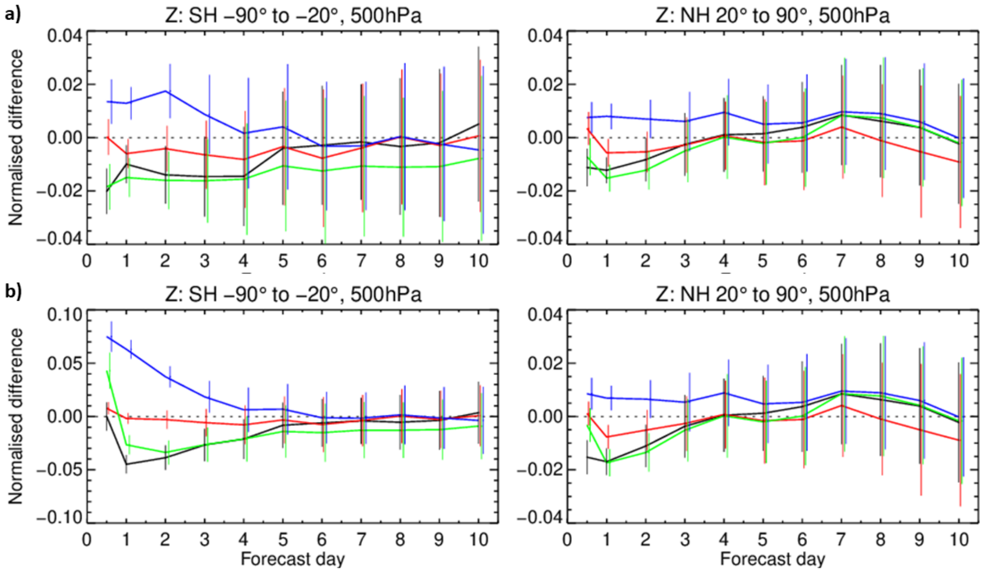

Figure 10: The impact of assimilating Spire (black), COSMIC-2 (red) and (Spire+COSMIC-2) (green) on Z500 RMSE scores. The control experiment includes all measurements used operationally. Removing all GNSS-RO from the control is also shown (blue). Verification is against own analysis (a) and against operational analysis (b), noting the different vertical scales.

4.3.1. Consistency of real and simulated COSMIC-2 observations in the EDA

One important new result is a comparison of the impact of real and simulated COSMIC-2 GNSS-RO measurements in the EDA system. The Harnish study established that similar numbers of real and simulated GNSS-RO data had a similar impact in the EDA, but the new results are a more systematic study. This is done by simulating GNSS-RO measurements at the observed COSMIC-2 times and locations, using the ECMWF operational analyses with the bending angle simulation code and the observation uncertainty model developed in the Harnisch study.

Figure 11 shows vertical profiles of the 12-hour forecast temperature spread reductions obtained with both real and simulated COSMIC-2 data. Overall, the agreement between the real and simulated data is very good, and we also see similar levels of consistency for geopotential, wind and relative humidity. These results suggest that the simplifications and approximations in the GNSS-RO simulation process (Section 4.1.3) do not have a significant impact on the EDA spread estimates, and that we are able to predict the performance of real GNSS-RO data in the EDA reasonably accurately.

There are some small differences in the tropics, with the simulated data giving larger spread reductions in the troposphere, and smaller reductions above 100 hPa. The differences in the troposphere are still under investigation, but they may be related to a lack of vertical error correlations in the simulated data. In the tropical stratosphere, we have found that the simulated (o-b) departures are too large when compared to real data because of an ad-hoc inflation of the observation errors introduced by Harnisch et al. (2013) to reproduce the large observed departure statistics near the tropical tropopause. When this factor is reduced or removed, and the consistency between the real and simulated departure statistics is improved, the spread values show much better agreement.

Overall, the results show that the simulated data provides a good approximation of how the real COSMIC-2 data changes the spread values in the current EDA system. It is also worth emphasising that the real COSMIC-2 measurements improve the estimate of the atmospheric state in the OSEs, but the simulated data does not, making the point that similar spread reductions do always not lead to the same forecast impact.

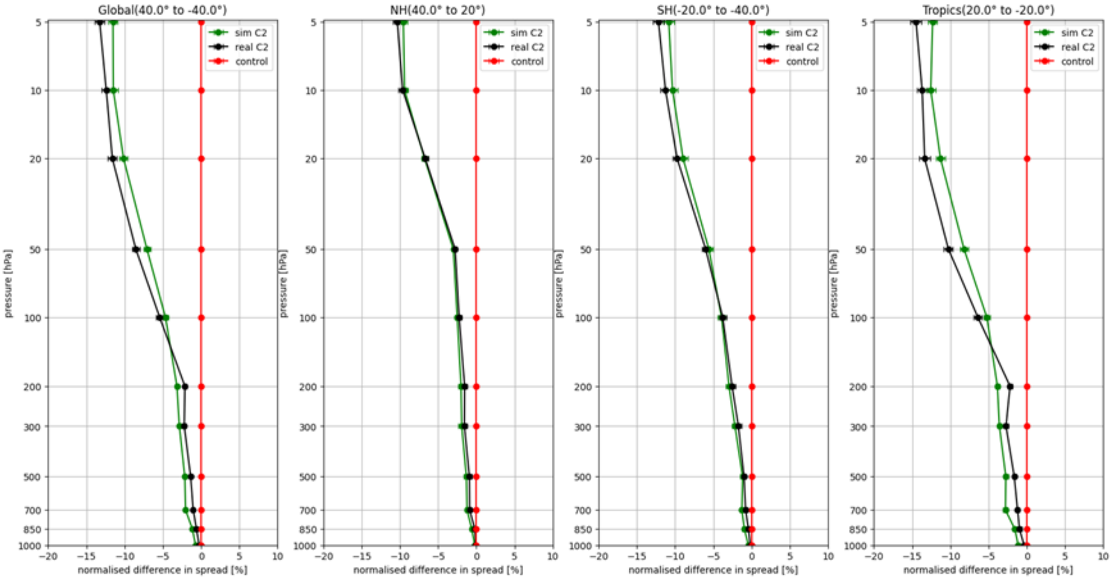

Figure 11: The vertical profile of 12 hour forecast temperature spread reductions, comparing the impact of real (black) and simulated (green) COSMIC-2 measurements in the EDA. The results are given as a percentage of the temperature spread of the control experiment. The spread is computed for the period January 10 to February 10, 2020. The comparisons are limited to ±40 latitude band sampled by the COSMIC-2 data. The control experiment includes GNSS-RO measurements used operationally in this period.

4.3.2. Relationship between EDA spread and radiosonde departure statistics

We have progressively added **real** GNSS-RO measurements in both OSEs and EDA simulations to investigate the relationship between the EDA spread changes and the OSE short-range forecast error statistics. Some of these results were presented to the SAC in 2021.

The change of radiosonde short-range forecast departure statistics computed from OSEs are compared with the corresponding "raw" EDA spread values. More specifically, ignoring correlations between the observation and short-range forecast errors, it was assumed that the change in variance of the radiosonde departures, $\sigma^2_{(o-b)}$, could be written as linear function of the EDA spread squared, s^2, using

$$\sigma^2_{(o-b)} = ms^2 + c$$

where ideally the intercept, c, should be related to the radiosonde observation error statistics. An example is shown in Figure 12, for the tropics at 100 hPa where the GNSS-RO information content is high.

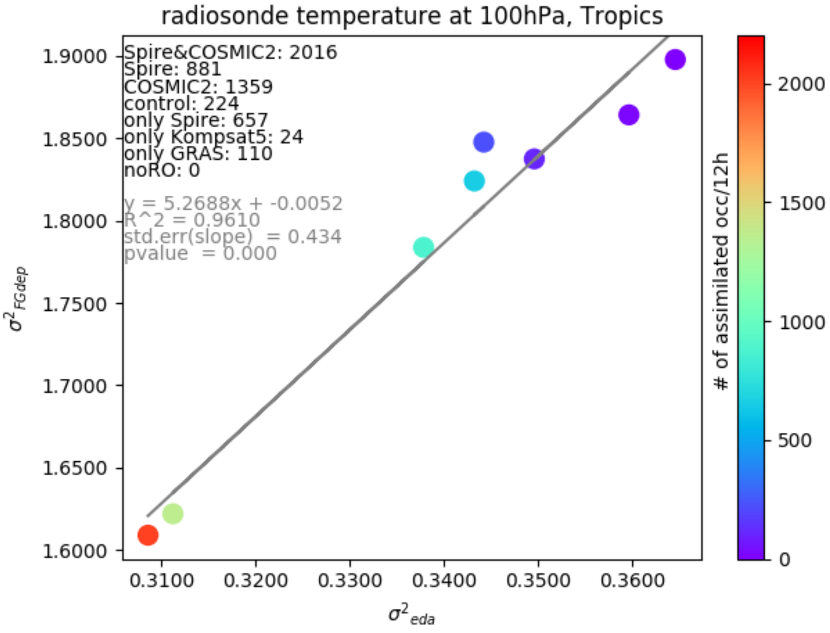

Figure 12: The variance of radiosonde short-range forecast temperature departures at 100 hPa in the tropics from the OSEs, versus EDA spread variance, as the number of GNSS-RO measurements is increased.

In general, the linear model is a good representation of the spread-skill relationship in the lower/middle stratosphere up to 10 hPa, but as noted by the SAC in 2021, the gradient values, m, tend to be too large, and in some cases the computed offset, c, can be negative, implying a negative radiosonde uncertainty value. The gradient values in the stratosphere are typically in the range m=5-11, depending on level and region. We obtain similar results when we compare the EDA spread values with the departure statistics of the EDA control member, so the resolution difference between the OSE and EDA does not account for this.

The EDA system is under-dispersive (Section 3.2), so it is expected that the absolute impact of the GNSS-RO will be underestimated. As discussed in Section 3.2, a global scaling factor of 1.34 is currently applied to derive background error standard deviations from the raw EDA spread, so introducing a factor of 1.8 (=1.34^2) in the m values would be consistent with this. However, this factor has been derived primarily to optimise 4D-Var forecast performance, rather than to calibrate spread/skill relationships for specific levels. The factors found here are more in line with results from an evaluation of spread/skill relationships for radiosonde temperature departures obtained by Bormann and Bonavita (2013, see their Figure 4), albeit using an older EDA system over a different period. In their study, they noted that different scaling factors were applicable to different observations, and scaling factors required for AMSU-A channels 8-9 were much lower than for radiosondes in the same vertical interval. One interpretation is that the EDA represents uncertainty at broader vertical scales more accurately than at the fine vertical scales, and this may be one explanation for the scaling factors observed here. However, this interpretation still requires further investigation.

The EDA only provides an estimate of the random error statistics, but the GNSS-RO measurements also have an additional role as anchor measurements within the NWP system. It is difficult to isolate the additional value of this anchoring in the EDA computations, but it may improve the radiosonde temperature departure statistics by correcting spatially varying biases. Figure 13 illustrates the size and spatial variability of the mean temperature analysis differences at 100 hPa, comparing all RO minus no RO. These changes in the mean state are large when compared to the changes in the EDA spread values.

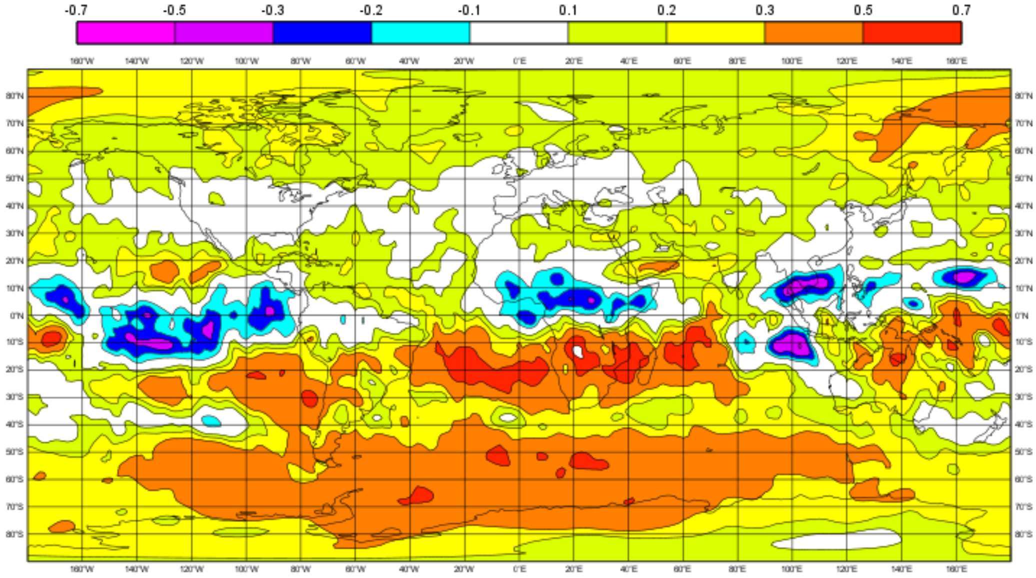

Figure 13: Mean temperature analysis difference (K) at 100 hPa. ALL RO, including COSMIC-2 and Spire, minus no RO.

4.3.3. Summary of recent GNSS-RO work

The assimilation of real COSMIC-2 and Spire observations has a clear positive impact. Importantly, we have found that real and simulated COSMIC-2 observations have a similar impact on the EDA spread. The improvements in radiosonde departure statistics indicate that the changes in the EDA spread underestimates the absolute impact of the GNSS-RO measurements in the lower/middle stratosphere. The global scaling factor used to inflate the raw EDA spread values for operational applications does not resolve this problem, and further work is required (See also Figure 18).

4.4 Constellations of small satellites carrying microwave sounders

The most recent application of the EDA-method is to estimate the expected impact from potential future constellations of small satellites carrying microwave (MW) sounding instruments, as part of an ESA-funded study (Lean et al., 2021a, b). The aim of such a constellation is to provide higher temporal sampling than currently available. This is expected to be beneficial to observe fast-evolving cloud and humidity features, but also to reduce the effective observation noise through higher measurement numbers. The continued benefit of adding MW sounding data has been demonstrated in observing system experiments with existing observations (Duncan et al., 2021), raising some expectation that further improvements in temporal sampling would lead to further benefits. Developments in satellite and sensor technology make it feasible to launch MW sounding instruments on small satellites or even cube-sats, thus making such constellations a possibility, as a complement to the backbone global observing system.

The study considered different orbit and instrument scenarios to probe two key aspects of the constellation design: how the impact varies with the number/ distribution of satellites, and the relative benefits of different sets of channels (183 GHz humidity-sounding only or with additional temperature sounding in the 50 GHz band). The constellations vary in size between 8 and 20 satellites, and the orbital planes are optimised to complement four existing MW sounders in the 9:30 (Metop) and 13:30 (JPSS) orbits (Table 1), part of the CGMS backbone constellation. Hypothetical instrument payloads were considered, with channels based on sub-sets of channels envisaged for the Arctic Weather Satellite.

Table 1: Satellite constellations with MW sounding considered. In the real-data cases, combinations of AMSU-A/MHS or ATMS were normally used, with the exception of the 5th orbit in the Metop/JPSS+ constellation, for which a combination of the NOAA-15 AMSU-A and the F-17 SSMIS was used (both in approximately a 6:30 orbit during the study period). For each constellation with MW sounding data, separate EDA experiments were run with humidity-sounding channels assimilated only, and with temperature and humidity-sounding channels assimilated.

(*) Strictly, 8 satellites with real data are used, though only 7 sets of temperature + humidity sounding capabilities.

Constellation name	Type of orbits	Number of orbital planes	Number of satellites
Real data			
No MW sounders	-	0	0
Metop/JPSS baseline	Sun-synchronous	2	4 (Metop-A/B; S-NPP, NOAA-20)
Metop/JPSS+	Sun-synchronous	5	8* (Metop-A/B; S-NPP, NOAA-15/ 18/19/20, F-17)
Simulated new data, added to the Metop/JPSS baseline with real data			
Polar	Sun-synchronous	4	8

Polar+	Sun-synchronous	7	14
Polar++	Sun-synchronous	10	20
4x2	Mid-inclination (60º)	4	8
6x2	Mid-inclination (60º)	6	12
Polar + 4x2	Sun-synchronous + mid-inclination (60º)	8	16

4.4.1. Observation-specific adaptations

The new observations were simulated and assimilated in the all-sky framework developed for MW-sounding radiances (Geer et al., 2014, Duncan et al., 2022), using RTTOV-SCATT as the observation operator (Geer et al., 2021). The overall approach of simulating the new observations largely followed that of the recent GNSS-RO work, that is, the observations were simulated from high resolution (TCo1279, 9km) ECMWF analysis trajectories. Over land and snow/sea-ice, where the assimilation of real sounding data relies on an emissivity retrieval from window-channel observations, alternative approaches to specify the surface emissivity needed to be used, as the emissivity retrieval is not possible. Here, values from an emissivity atlas were employed over snow-free land, and typical values over snow and sea-ice surfaces. The latter are relatively crude assumptions, reflecting the current lack of sufficiently accurate physically-based surface emissivity modelling.

An important new aspect to address during the simulation of the cloud-affected brightness temperatures was how to treat representation error. For all-sky assimilation of MW radiances, this is considered dominant in cloudy situations, and this aspect is reflected in the situation-dependent observation error model used during the assimilation (Geer and Bauer, 2011). The observation error model assigns larger values in cloudy regions, dependent on a cloud indicator which characterises the presence of clouds in the observations or the background. Applying a similar model during the simulation of the new observations (prior to the EDA) was considered. However, during the course of the work it was found that adding random perturbations according to the instrument noise alone led to 4D-Var background departure statistics that were remarkably consistent with those of real observations. An example of this is given in Figure 14 which shows standard deviations of background departures as a function of the cloud indicator used in the observation error model. The good agreement is likely the result of several factors. One is chaotic error growth relating to clouds. Another is differences in the spatial resolution, as the brightness temperatures were produced using TCo1279 (9km) trajectories, but assimilated at lower resolution (TCo399, ~25 km, as in the EDA configuration), hence introducing some form of spatial representation error. Differences in the moist physics may also play a role, with the old physics parametrization used in the observation simulations, whereas the new moist physics (Bechtold et al., 2020) were used in the assimilation, hence introducing a representation error relating to the cloud-representation. Given these results, it was decided to only add random perturbations consistent with the specified instrument noise when producing the simulated observations. Note, however, that the perturbations subsequently added in the EDA were nevertheless based on the full all-sky observation error model, with larger values in cloudy regions. The parameters for this error model were chosen in line with values used in the assimilation of real data.

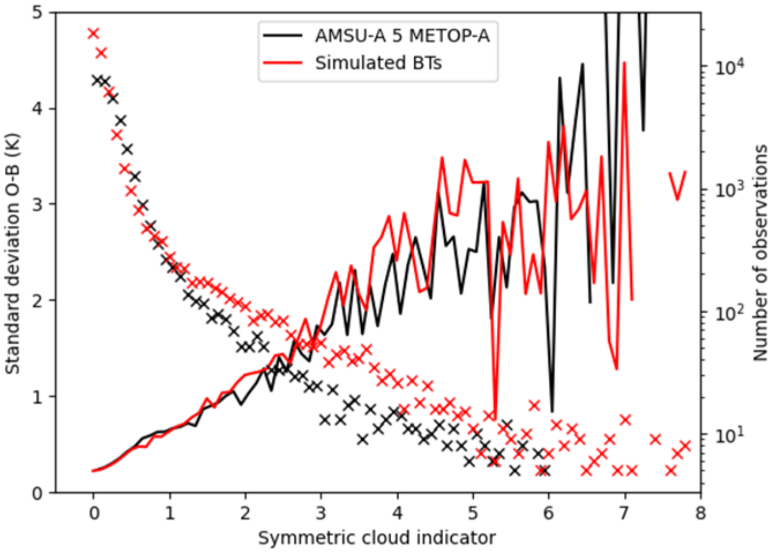

Figure 14: Increase in standard deviation of O-B (solid lines) with increasing presence of cloud signal for real Metop-A AMSU-A channel 5 data (black) and simulated small-satellite data for the equivalent channel (red). Low values of the cloud indicator indicate clear-sky regions, whereas increasing values indicate increasing presence of cloud signals. Crosses indicate the number of observations (right x-axis). Data are from the period 9-14 June 2018, over land surface only and from latitudes between 60ºN-60ºS.

To test the sensitivity to the perturbations applied during the observation simulation, an additional EDA experiment using the Polar (8 satellite) constellation without these initial perturbations was also run. The observation error modelling applied in the EDA (and hence the perturbations applied in the EDA) were unchanged. Figure 15 shows a representative example that the effect of the initial perturbations applied has very little impact on the overall changes in EDA spread compared to the overall signal. It gives some further indication that the modelling of the perturbations applied during the simulation of the observations is a secondary influence on the EDA spread, at least as long as these perturbations are smaller than the perturbations subsequently added in the EDA.

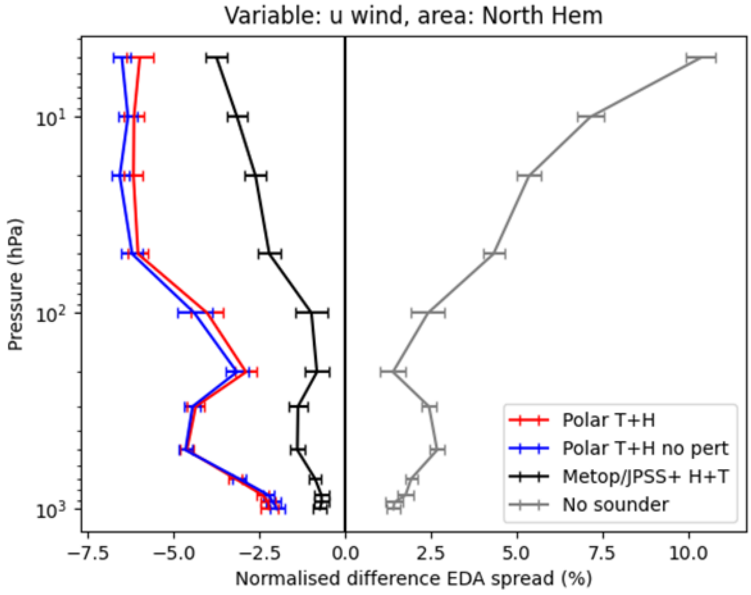

Figure 15: Vertical profile of normalised percentage reduction in EDA spread for the U wind component in the northern hemisphere for experiments where the simulated data have been assimilated with and without initial perturbations (red and blue lines respectively) or real MW data have been added or denied (black and grey lines respectively) compared to a baseline with 4 MW sounders.

One aspect not previously encountered in either the Aeolus or the GNSS-RO work was the question of how to deal with observational biases for the simulated data. Observational biases are a common feature in the assimilation of radiance observations, arising, for instance, from systematic errors in the calibration or the observation operator. For real data, these are removed during the assimilation using Variational Bias Correction (VarBC). For the simulated data, we made no attempt at modelling such observational biases. This was a pragmatic decision, motivated by the fact that any modelling of bias structures would be highly speculative and subject to the maturity of the level-1 processing of the new data. To avoid that the EDA treats the new observations as anchors for the bias correction of other observations, we however activated VarBC for the new observations, so required the system to estimate bias correction parameters. Corrections to the small satellite data remained stable and small in magnitude compared to those applied to real MW data as expected. The approach used is considered equivalent to assuming that there are no significant biases that are systematic for all satellites across the constellation, and that inter-satellite biases for a given channel average out to near zero when all satellites are considered together. In terms of global biases, this is consistent with the current experience from, for instance, AMSU-A, MHS, or ATMS data. The approach is also justified by assuming that the constellations considered complement a backbone observing system of MW sounders which could be used for inter-calibration exercises. Nevertheless, it remains an open question to what extent observational biases and their treatment through VarBC affect the EDA spread - an aspect that would be worth investigating further in the future.

4.4.2. EDA spread results for different constellation choices

Figure 16 illustrates some of the key results in terms of the EDA spread reduction for the constellations considered as a function of the number of sounding locations assimilated. Reductions are shown relative to a system in which all MW sounding data were denied, but otherwise the full real observing system was used. The data point furthest to the left represents the Metop/JPSS baseline using real data. The next two datapoints depict the spread reduction resulting from adding three existing MW sounders to this baseline (Metop/JPSS+), either with humidity-sounding capabilities only (red) or with temperature and humidity sounding (black). The remaining datapoints show spread reductions from adding instead simulated data from the potential future constellations considered, again either with humidity-sounding (red) or temperature- and humidity-sounding capabilities (black). The trend of the data points from adding simulated data extends smoothly from those using only real observations. This is an important and very reassuring result which adds further confidence in the simulations.

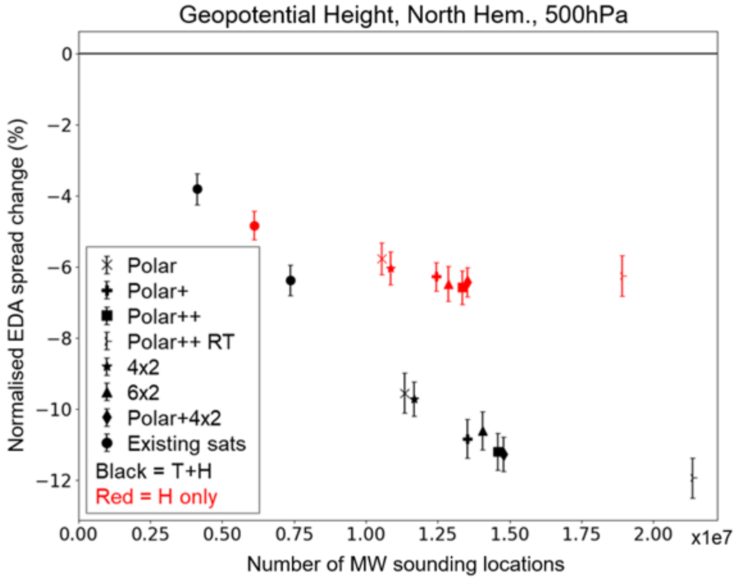

Figure 16: Percentage EDA spread reductions with increasing observation numbers for geopotential height at 500 hPa from a baseline with no MW sounder data for all constellation options and baselines. Black and red symbols indicate respectively the addition of temperature and humidity sounding channels or humidity channels only to the Metop/JPSS baseline. Different symbols denote the different simulated data scenarios or the use of real MW data where the additional "Polar++ RT" refers to the reduced thinning sensitivity experiment (see main text). Data are from the Northern Hemisphere (latitude > 20°) over the period 8-28 June 2018.

Some key points are readily apparent from Figure 16, providing relevant guidance for the constellation design:

1. There is continued benefit from adding MW sounding data, within the range covered, and even the smallest constellation considered ("Polar", ie 8 satellites in 4 orbital planes) brings sizeable

benefits, more than doubling the impact of MW sounders compared to the Metop/JPSS baseline. As we know from OSEs, this is a very significant impact (Duncan et al., 2021).

2. There is a clear added benefit from the temperature sounding channels (compare red and black data points), highlighting the value of a more complex instrument that includes such capabilities. The benefit from temperature-sounding capabilities most likely results from the effective noise reduction due to repetitive measurements: for MW temperature-sounding, the size of typical errors in the background is comparable to that of the instrument noise of the available observations, hence reducing the effective noise is especially beneficial.

3. As in the case of the GNSS-RO data, the rate of the spread reduction decreases for larger constellations, with the shape of the curve broadly similar to that seen for the GNSS-RO data.

The EDA also captures that the assimilation of MW-sounding observations leads to improvements for wind forecasts, through the ability of 4D-Var to infer information on the dynamics from observations primarily sensitive to temperature, humidity, and clouds. This is highlighted in Figure 17, which shows the EDA spread reduction for the Northern Hemisphere and the Tropics for some selected constellations. The relative benefit for wind of adding temperature-sounding channels to the humidity-sounding channels differs between the two regions shown (compare, for instance, the purple solid and dashed lines). The different importance is consistent with the primary mechanisms considered relevant for obtaining wind information: over the Tropics, the humidity-sounding channels play a larger role, via the 4D-Var humidity/cloud tracing, whereas over the extra-tropics there is more significant benefit from the temperature-sounding channels, as geostrophic balance becomes a leading mechanism. The real data show a similar behaviour (compare the solid black and dotted lines), and the finding is also consistent with experiences from OSEs. It is reassuring that the EDA captures these different behaviours.

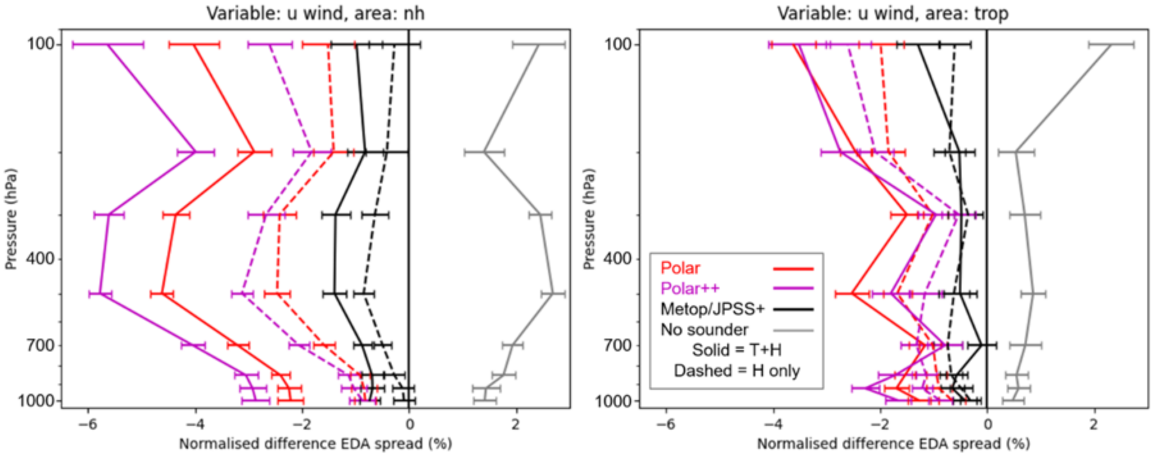

Figure 17: Change in the EDA spread relative to the Metop/JPP constellations for the u-component of wind over the Northern Hemisphere extra-tropics (left) and the Tropics (right) for the constellations listed in the legend. Dashed lines indicate experiments with humidity-sounding only, whereas solid lines indicate combined use of temperature and humidity-sounding.

One rather practical aspect encountered during the study was the role of thinning applied prior to the assimilation of MW-sounding data. In line with current practice for real data, observations for the simulated data were thinned spatially, selecting only one observation within a 110 km distance per half-hourly time-slots. This is to limit the effect of spatially correlated observation errors which we can currently not account for during the assimilation. In addition, data from all satellites of the hypothetical future constellation were thinned together (following the AMSU-A approach), resulting in many observations being thinned out at higher latitudes where there is most overlap. The thinning practice is the reason why the increase in the number of data points shown in Figure 16 for the different constellations is not as large as might be expected given the increase in the number of satellites. The effect of this was particularly severe as the phasing of the satellites in the small-satellite constellations was not optimised to avoid this.

To investigate the role of thinning all satellites of the constellation together, we also ran a pair of EDA sensitivity experiments in which we assimilated the largest, Polar++ constellation with the satellites spatially thinned separately rather than all together. This most extreme point in the number of observations in Figure 16 suggests that it is possible to approach a point where the benefit from additional measurements slows considerably (see “Polar++ RT”). Note, however, that the additional data here comes from observing a similar location at a similar time, and hence the only mechanism to provide further information is effective noise reduction, rather than the observation of temporally evolving structures. This likely explains at least partly the lack of further spread decrease for the humidity-only scenario, as instrument noise is not considered a limiting factor for these observations. The comparison also highlights the sensitivity of the EDA results to practical assimilation choices and assumptions on error characteristics, an important element to keep in mind when interpreting the results.

4.4.3. Comparing EDA spread reductions and OSE results

The four EDA experiments with real observations also allowed a comparison of the EDA spread reductions and forecast error reductions in a similar way as presented in Section 4.3.2 for the GNSS-RO work. The measure of short-range forecast error was again the variance of background departures for radiosondes. As for the GNSS-RO results, in general there is a trend for increasing variance of the radiosonde O-B with variance of the EDA spread such as in the examples shown in Figure 18. While the low number of data points limits interpretation, many levels/regions show a broadly linear relationship, though this is less clear for the troposphere in the tropics, where the signal in the EDA spread reduction is weaker. The broadly linear relationship again highlights that there is a reasonable agreement between the EDA and the OSEs in these areas, at least for relative forecast impacts. Nevertheless, the slope and intercepts of the regression lines vary between pressure levels and variables within the MW study, with slopes for temperature in the range of 1.1-2.2 over the Northern Hemisphere extra-tropics and 1-7 elsewhere. Results for wind are broadly similar. As discussed in Section 4.3.2, the recognised under-spread of the EDA is likely a factor why the forecast error reduction measured by radiosondes is larger than the EDA spread reduction, as indicated by slope-values larger than one.

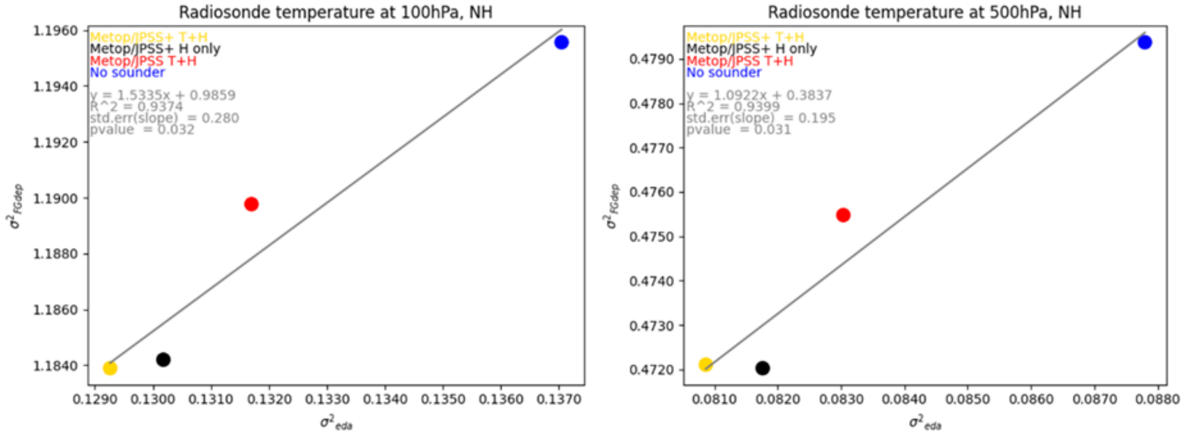

Figure 18: Scatter plot of the variance of the radiosonde (O-B) and variance of the EDA spread for temperature at 100 hPa (left) and 500 hPa, both Northern Hemisphere (right).

The slope values have also been compared to those obtained from the GNSS-RO study, and the values differ considerably, with the GNSS-RO scaling factors typically larger than the ones for the MW. However, the two studies were conducted over different seasons and years (June 2018 for the MW work, and January-March, 2020 or the GNSS-RO work), so a direct comparison of the scaling factors is questionable, as at least some of the differences likely reflect seasonal and inter-annual differences in the performance of the EDA. In line with seasonal aspects playing a particular role, profiles of EDA spread for the two studies differed strongly in size for the extra-tropics when comparing the same hemispheres, but they were broadly similar when comparing spread values for opposite hemispheres. Values for the variance of radiosonde departures also showed better agreement for the opposite hemispheres. The relevance of the scaling factor and any similarities and differences for different observing systems nevertheless warrants further research. One aspect currently being considered is whether the spatial sampling of the different observing systems also influences these gradient values.

4.5 Medium-range impact assessment with the EDA

Ideally, we would like to provide an estimate of the observation impact at medium-range, but the EDA studies usually focus on spread reductions achieved at the short-range where the observation signal is clearest. In general, if we reduce the errors at short-range, we are likely to improve the medium-range errors, but this is only considered a partial predictor of medium-range skill. We encounter a similar situation for the interpretation of FSOI, which also only estimates short-range forecast skill. In principle, we can extend the EDA spread calculations to longer ranges (without adding the singular vectors used in the ENS system) as part of the assessment studies. For example, Figure 19 compares the EDA temperature spread reductions achieved with real and simulated COSMIC-2 measurements at day-5. There is still reasonable consistency between the real and simulated data which is encouraging, and as expected the COSMIC-2 spread reductions are smaller than achieved at 12 hours (Figure 11).

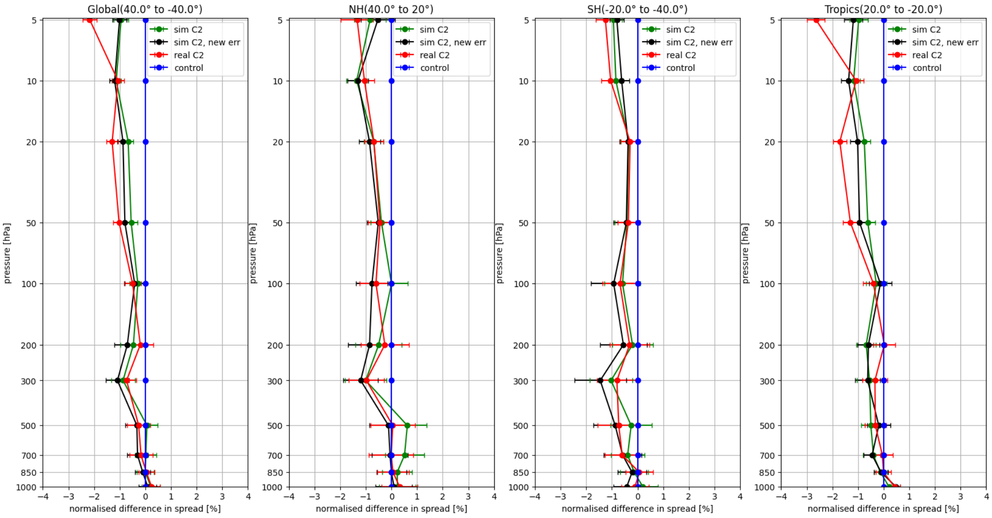

Figure 19: As Figure 11 but showing day-5 EDA temperature spread reductions for real (red line) and simulated (green line) COSMIC-2 measurements. This includes a revised error model for simulated data (black line), which gives better consistency with the real data for 12 hr spread values.

An open question is whether the day-5 spread estimates provide useful additional information, given that we know the EDA alone does not fully characterise the forecast error statistics in the medium-range, and that singular vector-based perturbations are needed to obtain a reliable ensemble forecast (Section 3.3). Also, to obtain statistically significant spread differences, we would probably need to run longer experiments, which would increase the cost of the EDA experimentation.

4.6 New/planned activities

Two new EDA studies started in June 2022, through a EUMETSAT-funded activity. In these studies, we will consider two potential future missions, namely the EPS-Sterna constellation and a future Doppler Wind Lidar as Aeolus follow-on. A new aspect is that both missions will be investigated in a consistent way, over the same period, using the same baseline observing system of real observations. This will allow cross-comparisons of the impact and investigations into the complementarity of these missions. We have not previously performed such comparisons of different observing systems with the EDA-method. Given the potential differences in scaling factors when translating EDA spread results into forecast error reductions noted earlier, such comparisons will require some care.

EPS-Sterna Constellation

The EPS-Sterna constellation consists of a fleet of small satellites with cross-track scanning MW instruments, with temperature and humidity sounding capabilities in the 50, 183, and 325 GHz bands. It is intended to complement the 3-orbit CGMS baseline, to provide better temporal sampling for MW sounding. The requirements for the constellation are currently being refined by EUMETSAT in collaboration with ESA, users and industry, for instance, in terms of the number of orbits covered, satellites flown, and expected instrument performance.

The EDA simulations to be conducted will build on and complement the work presented in Section 4.4, and results are expected to contribute to the required trade-off analyses. A particularly novel aspect will be estimating the impact of the new 325 GHz channels which are similar to channels to be flown on the Ice Cloud Imager on EPS-SG, but are not currently available from space.

Future Doppler Wind Lidar

The Aeolus Doppler Wind Lidar (DWL) did not meet the pre-launch mission requirements, but the horizontal line of site (HLOS) winds have had a significant positive impact in NWP systems (Rennie et al., 2021).

An operational follow-on DWL mission is being considered, and Phase 0 of preparation activities started at ESA and EUMETSAT in 2020. Consolidated requirements are also currently being compiled. It seems reasonable to expect that a future DWL mission will improve upon the Aeolus performance in terms of HLOS quality. Therefore, it is useful to estimate how this anticipated improvement will translate into an additional DWL impact in an NWP system. The EDA approach will be used to test this. One simulated DWL dataset will be produced with noise characteristics consistent with the actual Aeolus performance, and the other will be based on the expected performance of the future DWL system.

Future MW-sounding systems from NOAA

In addition, a further EDA-activity will be done in collaboration with NOAA and the Cooperative Institute for Satellite Earth System Studies (CISESS). The aim will be to evaluate the NWP benefit from a future MW sounder that is being considered by NOAA as a follow-on to the ATMS instrument, with added channel capabilities in the 118 GHz band. The study will consider the impact of such an instrument in the 13:30 and 5:30 orbits, and relate this impact to that of the MW sounders on the current heritage POES satellites as well as that of ATMS.

5. Summary

This paper has reviewed and summarised approaches used at ECMWF to evaluate the benefit of observations in global NWP. The main emphasis has been on evaluating the expected impact of future observations for which ECMWF has established a 4D-Var information-content approach that builds on the operational EDA system. However, we have also reviewed methodological aspects of the use of OSEs and FSOI for measuring impact of existing observations, as well as key features of the performance of the ECMWF EDA.

The global observing system is evolving, and OSEs are considered the gold-standard for evaluating the present impact of real observations. ECMWF complements this with FSOI evaluations for diagnostic purposes. Recent OSEs have highlighted challenges in terms of reliable verification of short-range forecast impact, and they exhibit a healthy robustness and resilience of the global system to denying single observing systems for NWP. Given the current state of the global observing system, it is increasingly important to optimise the complementarity of different future systems and to evaluate opportunities offered by new technology. OSEs and FSOI can provide important insights into how the global observing system should evolve, but simulations of the future impact of specific observing systems are an important ingredient for achieving such an optimised design.

The EDA method for assessing new observations provides a theoretical estimate of the expected reduction in analysis and short-range forecast uncertainty, as a result of assimilating the new data. This information is usually averaged over large spatial scales and for periods of order a month, so the EDA results should be interpreted statistically. The EDA is less useful for quantifying the impact of the new data on specific weather events or case studies. The method is similar to a 1D-Var information content study, but it is more general, because it is a cycling DA system which includes the forecast model, so we are also able to estimate the impact on variables that are indirectly related to the measured quantities (e.g., wind impact from MW sounders). The current operational EDA is known to be under-dispersive, and this means that the absolute impact of new observations is likely to be underestimated. However, future assessments of new observations will automatically benefit from improvements in the operational EDA system. In addition, it is usually argued that we are comparing the impact of different observation scenarios in a given set of the EDA experiments, and the relative impact is important.

The main results from the early Aeolus (Section 4.2.1) and GNSS-RO (Section 4.2.2) EDA studies are reasonably consistent with the subsequent achievements with real data. Aeolus has a good forecast impact for a single instrument, for example, improving tropical stratospheric winds out to day-9 (100 hPa). The GNSS-RO impact on Z500 RMSE scores at short-range roughly doubled when testing COSMIC-2 and Spire (Figure 10). However, direct comparisons between the EDA spread reductions and improved forecast scores at short-range are sensitive to the choice of verifying analyses, as discussed in Section 2.1.1. We have also found that some of the percentage improvements in EDA spread given in the Harnisch study are not achievable today, because the global observing system has improved considerably since 2008, and the baseline “noRO” spread values are now much smaller. On the key question of saturation of GNSS-RO impact, Privé et al. (2022) did not see saturation when testing up to 100,000 occultations per day in OSSEs, supporting the EDA work.

The impact of real and simulated measurements in the EDA system has been tested and the results are very encouraging from a methodology viewpoint. The direct comparison of real and simulated COSMIC-2 measurements in the EDA (Figure 11) suggests we can predict the impact of new GNSS-RO data in the EDA accurately with the simulation code developed for the Harnisch study. In addition, the EDA response to real and simulated microwave data appears to be consistent (Figure 16). Overall, if a simulated dataset has realistic (o-b)s, and it is assigned a reasonable **R** matrix, these results suggest that the spread reductions will be similar to what will be achieved with real measurements with the same observation error characteristics.

There are some outstanding questions that require further work. Firstly, it is clear that changes in the EDA spread values can underestimate the actual performance of the observations, particularly for GNSS-RO in the stratosphere (Figure 12). This is to be expected in general, because the EDA is under-dispersive. Areas being investigated to improve the reliability of the EDA include (Section 3.2): improved model uncertainty parameterisations; better modelling of observation errors including horizontal correlations; reduction of the resolution gap between the EDA and the high-resolution analysis.

The different scaling factors required for GNSS-RO and microwave data (Figure 18), albeit for different seasons, requires investigation, particularly if we wish to compare different observing systems with the EDA in the future. The role of anchor measurements and bias correction may be relevant here. Real GNSS-RO measurements anchor the system because they can be assimilated without bias correction, but this key measurement characteristic is not captured in the EDA assessment. In contrast, the microwave measurements will require bias correction and they are assimilated with VarBC so that they are not treated as anchor measurements, but biases are not currently applied in the simulation process.

The present EDA studies focus on changes in the spread at short-range, and interpret this as a partial predictor of medium-range forecast impact. The EDA experiments can provide an estimate of medium-range impact, and we have found that simulated and real COSMIC-2 measurements produce similar spread reductions at day-5 which is encouraging (Figure 19). However, the interpretation of such results is complicated by the fact that singular vectors are required to produce a reliable ensemble at medium-range. For this reason, we consider short-range results to be most robust. The emphasis on short-range impact is in common with the use of FSOI, which also is strictly only applicable to characterise short-range impact.

In principle, well calibrated OSSEs potentially provide more information on the medium-range impact of new observing systems, but developing and maintaining a well calibrated OSSE system, which also keeps pace with the evolving global observing system, is a major challenge. It is probably easier to test the impact of new observations on top of the latest global observing system with the EDA, because it is used in operations and it will include the latest observations. However, we do not see the EDA and OSSEs as in direct competition. It is important to recognise that the impact of new observations will vary depending on the details of the implementation in each NWP system. Having access to both OSSEs and EDA estimations from different systems (and possibly other methods, like Sensitivity Observing System Experiments, Marseille et al., 2008) should help isolate system dependent aspects from the real impact, and provide better overall guidance to the agencies. The recent OSSE results from Privé et al. (2022) illustrate this point. Adding GNSS-RO clearly degrades winds in the tropical stratosphere in that system (their Figure 16), but this not case in either the EDA experiments or OSEs with real GNSS-RO, so this apparent degradation should be ignored. This kind of problem is easy to spot for an existing observing system where OSEs are available, but it will be much harder for completely new observation types, so multiple approaches should be useful. In this context, we note that the assessments of the future DWL system coordinated by EUMETSAT, will include both ECMWF EDA estimates with OSSEs performed with the NOAA system. Similarly, an OSSE will be performed by Météo-France for the EPS-Sterna constellation, in addition to the EDA experimentation planned at ECMWF.

To date, we have used the EDA to assess potentially large changes to the global observing system, but usually involving just one observation type which is already used operationally and well understood. Going forward, the impact of different observing systems will be compared with the EDA, for instance in the context of the DWL and EPS-Sterna constellation work mentioned earlier. Doing so will implicitly assume that the EDA response reflects the short-range forecast uncertainty in a similar way for the two observing systems considered. As noted above, this aspect requires further investigation, for instance by conducting EDA experimentation with real data for the different observing systems over the same period and in the same system. At times, space agencies are also interested in more subtle questions for specific instrument design, such as how changes in horizontal resolution or sampling for a MW instrument affects the results, or what benefit the addition of a specific channel brings. Such questions are very difficult to answer with the EDA method (or indeed OSSEs), either as our modelling and assimilation of the observations is not sophisticated enough to assess such details, or as the signals are likely to be small, and uncertainties in the EDA method may dominate. We consider the EDA method to be best suited for relatively sizeable changes to the observing system.

In summary, although there are aspects that require further research, we believe the EDA method can provide useful information on the impact of future observing systems, that can complement the information provided by other techniques such as OSSEs.

6. Appendix: Measuring analysis quality and forecast skill

It is worth explaining some of the fundamentals of forecast verification as they help to explain many of the issues encountered when interpreting OSEs (see also ECMWF, 2018, Appendix A).

A typical measure of forecast skill is the square root of the mean squared error (RMSE). For a forecast or analysis f_i and a reference a_i, the RMSE is the square root of the MSE:

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}(f_i - a_i)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - a_i)^2}$$

(Eq. A1)

The index i represents one of n samples. A global estimate of the error in the 500hPa geopotential height, for example, would use a sample encompassing all latitudes and longitudes and different forecast base times.

The RMSE can be decomposed in a number of informative ways. First is the decomposition into the quadratic sum of the error standard deviation S and the mean error M:

$$\mathrm{RMSE} = \sqrt{S^2 + M^2}$$

$$S = \frac{1}{n}\sum_{i=1}^{n}(f_i - a_i - \mathrm{M})^2$$

$$\mathrm{M} = \frac{1}{n}\sum_{i=1}^{n}(f_i - a_i)$$

(Eq. A2)

This shows that RMSE comes from a quadratic sum of what could be termed transient and systematic errors.

However, the reference is not the truth, and the characteristics of the reference have an important bearing on the RMSE. This is seen by introducing the non-observable true state of the atmosphere t_i and decomposing the RMSE into the true error in the forecast, the true error in the reference, and allowing for the fact that the two errors may be correlated.

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}(f_i - t_i) + MSE(a_i - t_i) - 2\mathrm{Cov}(f_i - t_i, a_i - t_i)}$$

(Eq. A3)

Here the error covariance Cov() follows standard definitions (e.g. Wilks, 2006). This decomposition shows how the errors in the reference, $MSE(a_i - t_i)$, are part of the RMSE. If these are too large, they can dominate the RMSE and it becomes hard to detect changes in what we ultimately want to know, the errors in the forecast, $\mathrm{MSE}(f_i - t_i)$. This may be one of the main problems in using observations as the reference for verifying changes in the short-range forecast, since the short-range forecast is often more accurate than any single observation. However, if there is any correlation between the errors in the reference and the errors in the forecast, then these will artificially reduce the RMSE through the covariance term. This is why it is problematic to use analyses to verify the forecast, since these errors can be strongly correlated. There is no perfect reference.

Another helpful decomposition is similar to the previous one, but instead subtracts a climatology c_i:

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}(f_i - c_i) + MSE(a_i - c_i) - 2\mathrm{Cov}(f_i - c_i, a_i - c_i)}$$

(Eq. A4)

Unlike the truth in the previous decomposition, the climatology can be calculated. In this case (see ECMWF, 2018, Appendix A) the covariance term is then considered a good measure of true forecast skill beyond what we already know about climate. Hence the covariance term is the basis of the anomaly correlation coefficient (ACC) which is favoured in some verification contexts. By contrast, the RMSE does not just include the covariance term, but also the two MSE terms, which in this case are seen to measure the variability (sometimes known as activity) in the forecast and in the reference. When the analysis is used as the reference, observing system experiments can be affected by this issue because a change in the observational usage will often change the variability in the analysis. Changes in variability are also a confounding factor in any change to the forecasting system that alters the smoothness of the fields, such as a model resolution upgrade.

Although decomposed measures of skill can help mitigate against possible confounding effects in measures of total error like RMSE, focusing too strongly on measures such as the error standard deviation or the ACC means that other components of the error may be ignored (respectively the mean or the variability). The great advantage of the RMSE is that all components of the error are being tracked; the problem is that it takes effort to understand the origins of any change in the RMSE, and to decide whether those changes are important or not.

7. References

Bechtold, P., R. Forbes and I. Sandu, S. Lang and M. Ahlgrimm, 2020: A major moist physics upgrade for the IFS. *ECMWF Newsletter*, 164, 24-32, doi:10.21957/3gt59vx1pb

Bonavita, M., E. Hólm, L. Isaksen, and M. Fisher, 2016: The evolution of the ECMWF hybrid data assimilation system. *Q. J. R. Meteorol. Soc.*,142, 287-303, doi:10.1002/qj.2652

Bormann, N. and M. Bonavita, 2013: Spread of the ensemble of data assimilations in radiance space. *ECMWF Technical Memorandum*, 708, 29pp, doi:10.21957/edrq57brh

Bormann, N., H. Lawrence and J. Farnan, 2019: Global observing system experiments in the ECMWF assimilation system. *ECMWF Technical Memorandum,* 839, 24pp, doi:10.21957/sr184iyz

Bouttier, F., and G. Kelly, 2001: Observing-system experiments in the ECMWF 4D-Var data assimilation system. *Q. J. R. Meteorol. Soc.*, 127,1469-1488, doi:10.1002/qj.49712757419

Buizza, R., M. Leutbecher and L. Isaksen, 2008: Potential use of an ensemble of analyses in the ECMWF Ensemble Prediction System. *Q. J. R. Meteorol. Soc.*, 134, 2051-2066, doi:10.1002/qj.346

Cardinali, C., 2009: Monitoring the observation impact on the short-range forecast. *Q. J. R. Meteorol. Soc.*, 135, 239-250, doi:10.1002/qj.366

Cardinali, C., 2018: Forecast sensitivity observation impact with an observation-only based objective function. *Q. J. R. Meteorol. Soc.,*144, 2089-2098, doi:10.1002/qj.3305

Courtier, P., J.-N. Thépaut and A. Hollingsworth, 1994: A strategy for operational implementation of 4D-Var, using an incremental approach. *Q. J. R. Meteorol. Soc.*, 120, 1367-1387, doi:10.1002/qj.49712051912

Duncan, D.I., N. Bormann and E.V. Hólm, 2021: On the addition of microwave sounders and numerical weather prediction skill. *Q. J. R. Meteorol. Soc.*, 147, 3703-3718, doi:10.1002/qj.4149

Duncan, D. I., N. Bormann, A. J. Geer and P. Weston, 2022: Assimilation of AMSU-A in all-sky conditions. *Mon. Wea. Rev.*, 150, 1023-1041, doi:10.1175/MWR-D-21-0273.1

ECMWF, 2018: *ECMWF Forecast User Guide.* doi:10.21957/m1cs7h

Errico, R. M. and N. C. Privé, 2018: Some general and fundamental requirements for designing observing system simulation experiments (OSSEs). *World Weather Research Programme*, 2018-8, 24pp, https://ntrs.nasa.gov/api/citations/20190025338/downloads/20190025338.pdf (accessed 20220811)

Eyre, J.R., 2021: Observation impact metrics in NWP: A theoretical study. Part I: Optimal systems. *Q. J. R. Meteorol. Soc.*, 147, 3180-3200, doi:10.1002/qj.4123

Geer, A.J., P. Bauer and P. Lopez, 2010: Direct 4D-Var assimilation of all-sky radiances. Part II: Assessment. *Q. J. R. Meteorol. Soc.*, 136,1886-1905, doi:10.1002/qj.681

Geer, A.J. and P. Bauer, 2010: Enhanced use of all-sky microwave observations sensitive to water vapour, cloud and precipitation. *ECMWF Technical Memorandum*, 620, 41pp, doi:10.21957/mi79jebka

Geer, A.J. and P. Bauer, 2011: Observation errors in all-sky data assimilation. *Q. J. R. Meteorol. Soc.*, 137, 2024-2037, doi:10.1002/qj.830

Geer, A.J., F. Baordo, N. Bormann and S.J. English, 2014: All-sky assimilation of microwave humidity sounders. *ECMWF Technical Memorandum*, 741, 57pp, doi:10.21957/obosmx154

Geer, A.J., 2016: Significance of changes in medium-range forecast scores. *Tellus A*, 68, 21pp, doi:10.3402/tellusa.v68.30229

Geer, A.J., F. Baordo, N. Bormann, P. Chambon, S.J. English, M. Kazumori, H. Lawrence, P. Lean, K. Lonitz and C. Lupu, 2017: The growing impact of satellite observations sensitive to humidity, cloud and precipitation. *Q. J. R. Meteorol. Soc.*, 143, 3189-3206, doi:10.1002/qj.3172

Geer, A. J., P. Bauer, K. Lonitz, V. Barlakas, P. Eriksson, J. Mendrok, A. Doherty, J. Hocking and P. Chambon, 2021: Bulk hydrometeor optical properties for microwave and sub-millimetre radiative transfer in RTTOV-SCATT v13.0. *Geosci. Model Dev.*, 14, 7497-7526, doi:10.5194/gmd-14-7497-2021

Harnisch, F., S.B. Healy, P. Bauer and S.J. English, 2013: Scaling of GNSS Radio Occultation impact with observation number using an Ensemble of Data Assimilations. *Mon. Wea. Rev.*, 141, 4395-4413, doi:10.1175/MWR-D-13-00098.1

Houtekamer, P.L. and H.L. Mitchell, 1998: Data assimilation using and ensemble Kalman filter technique. *Mon. Wea. Rev.*, 126, 796-811, doi:10.1175/1520-0493(1998)126<0796:DAUAEK>2.0.CO;2

Houtekamer, P.L. and F. Zhang, 2016: Review of the ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.*, 144, 4489-4532, doi:10.1175/MWR-D-15-0440.1

Ingleby, B., B. Candy, J. Eyre, T. Haiden, C. Hill, L. Isaksen, D. Kleist, F. Smith, P. Steinle, S. Taylor, W. Tennant and C. Tingwell, 2021: The impact of COVID-19 on weather forecasts: a balanced view. *Geophys. Res. Lett.*, 48, 10pp, doi:10.1029/2020GL090699

Isaksen, L., M. Bonavita, R. Buizza, M. Fisher, J. Haseler, M. Leutbecher. and L. Raynaud, 2010: Ensemble of data assimilations at ECMWF. *ECMWF Technical Memorandum*, 636, 46pp, doi:10.21957/obke4k60

Janisková, M. and P. Lopez, 2013: Linearized Physics for Data Assimilation at ECMWF. *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. II)*, 251-286, doi:10.1007/978-3-642-35088-7_11

Kelly, G.A., P. Bauer, A.J. Geer, P. Lopez and J.-N. Thépaut, 2008: Impact of SSM/I observations related to moisture, clouds, and precipitation on global NWP forecast skill. *Mon. Wea. Rev.*, 136, 2713-2726, doi:10.1175/2007MWR2292.1

Lang, S., E. Hólm, M. Bonavita and Y. Trémolet, 2019: A 50-member ensemble of data assimilations. *ECMWF Newsletter*, 158, 27-29, doi:10.21957/nb251xc4sl

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

Lawrence, H., N. Bormann, I. Sandu, J. Day, J. Farnan and P. Bauer, 2019: Use and impact of Arctic observations in the ECMWF Numerical Weather Prediction system. *Q. J. R. Meteorol. Soc.*, 145, 3432-3454, doi:10.1002/qj.3628

Lean, K., N. Bormann and S.B. Healy, 2021a: Calibration of EDA spread and adaptation of the observation error model. *ESA Contract Report* for 4000130590/20/NL/IA, 23pp, doi:10.21957/1auh0nztg

Lean, K., N. Bormann and S.B. Healy, 2021b: Developing a flexible system to simulate and assimilate small satellite data. *ESA Contract Report* for 4000130590/20/NL/IA, 18pp, doi:10.21957/kjmxyh9xy

Leutbecher, M., S.-J. Lock, P. Ollinaho, S.T.K. Lang, G. Balsamo, P. Bechtold, M. Bonavita, H.M. Christensen, M. Diamantakis, E. Dutra, S. English, M. Fisher, R.M. Forbes, J. Goddard, T. Haiden, R.J. Hogan, S. Juricke, H. Lawrence, D. MacLeod, L. Magnusson, S. Malardel, S. Massart, I. Sandu, P.K. Smolarkiewicz, A. Subramanian, F. Vitart, N. Wedi and A. Weisheimer, 2017: Stochastic representations of model uncertainties at ECMWF: State of the art and future vision. *Q. J. R. Meteorol. Soc.*, 143, 2315-2339, doi:10.1002/qj.3094

Liu, Z.-Q. and F. Rabier, 2002: The interaction between model resolution, observation resolution and observation density in data assimilation: A one-dimensional study. *Q. J. R. Meteorol. Soc.*, 128, 1367-1386, doi:10.1256/003590002320373337

Lock, S-J, S. T. K. Lang, M. Leutbecher, R. J. Hogan and F. Vitart, 2019: Treatment of model uncertainty from radiation by the Stochastically Perturbed Parametrization Tendencies (SPPT) scheme and associated revisions in the ECMWF ensembles. *Q. J. R. Meteorol. Soc.*, 145, 75- 89, doi:10.1002/qj.3570

Lonitz, K., C. Marquardt, N. Bowler and S. Healy, 2021: Final Technical Note of 'Impact assessment of commercial GNSS-RO data'. *ESA Contract Report* for 4000131086/20/NL/FF/a, 72pp, doi:10.21957/wrh6voyyi

Lorenc, A.C. and R. T. Marriott, 2014: Forecast sensitivity to observations in the Met Office Global numerical weather prediction system. *Q. J. R. Meteorol. Soc.*, 140, 209-224, doi:10.1002/qj.2122

Lupu, C., C. Cardinali and A.P. McNally, 2015: Adjoint-based forecast sensitivity applied to observation-error variance tuning. *Q. J. R. Meteorol. Soc.*, 141, 3157-3165, doi:10.1002/qj.2599

Marquet, P., J.-F. Mahfouf and D. Holdaway, 2020: Definition of the Moist-Air Exergy Norm: a comparison with existing “moist energy norms”. *Mon. Wea. Rev.*, 148, 907–928, doi:10.1175/MWR-D-19-0081.1

Marseille, G.-J., A. Stoffelen and J. Barkmeijer, 2008: Sensitivity Observing System Experiment (SOSE)—a new effective NWP-based tool in designing the global observing system. *Tellus A*, 60, 216-233, doi:10.1111/j.1600-0870.2007.00288.x

Poli, P., S. Healy, F. Rabier, and J. Pailleux, 2008: Preliminary assessment of the scalability of GPS radio occultations impact in numerical weather prediction. *Geophys. Res. Lett.*,35, L23811, 5pp, doi:10.1029/2008GL035873

Privé, N.C., R.M. Errico and A.E. Akkraoui, 2022: Investigation of the potential saturation of information from Global Navigation Satellite System Radio Occultation observations with an Observing System Simulation Experiment. *Mon. Wea. Rev.*, 150, 1293-1316, doi:10.1175/MWR-D-21-0230.1

Rabier, F., E. Klinker, P. Courtier and A. Hollingsworth, 1996: Sensitivity of forecast errors to initial conditions. *Q. J. R. Meteorol. Soc.*, 122, 121-150, doi:10.1002/qj.49712252906

Rabier, F., H. Järvinen, E. Klinker, J.-F. Mahfouf and A. Simmons, 2000: The ECMWF operational implementation of four-dimensional variational assimilation. Part I: Experimental results with simplified physics. *Q. J. R. Meteorol. Soc.*, 126, 1143–1170, doi:10.1002/qj.49712656415

Rennie, M.P., L. Isaksen, F. Weiler, J. de Kloe, T. Kanitz and O. Reitebuch, 2021: The impact of Aeolus wind retrievals on ECMWF global weather forecasts. *Q. J. R. Meteorol. Soc.*, 147, 3555-3586, doi:10.1002/qj.4142

Rodgers, C. D., 2000: *Inverse Methods for Atmospheric Sounding.* 256pp, World Scientific Publishing, Singapore, doi:10.1142/3171

Tan, D.G.H., E. Andersson, M. Fisher and L. Isaksen, 2007: Observing-system impact assessment using a data assimilation ensemble technique: application to the ADM–Aeolus wind profiling mission. *Q. J. R. Meteorol. Soc.*, 133, 381-390, doi:10.1002/qj.43

Todling, R., 2013: Comparing two approaches for assessing observation impact. *Mon. Wea. Rev.*, 141, 1484-1505, doi:10.1175/MWR-D-12-00100.1

Wilks, D.S., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd Ed. Academic Press, Burlington, MA, USA.

WMO, 2020: *The 7th WMO Workshop on the Impact of Various Observing Systems on NWP*. https://community.wmo.int/meetings/NWP-7 (accessed on 20220811)