Technical Memo

European Centre for Medium-Range Weather Forecasts

937

A comparison of observing system experiments with the physics-based IFS and the data-driven AIFS

Niels Bormann, Ewan Pinnington, Gareth Jones, Matthew Chantry and Tony McNally

Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our web site under:
http://www.ecmwf.int/en/publications/

Contact: library@ecmwf.int

Abstract

This study intercompares results from Observing System Experiments (OSEs) with the physics-based Integrated Forecast System (IFS) and the machine-learning based Artificial Intelligence Forecast System (AIFS), with a focus on the performance in the troposphere. The OSEs characterise the respective denial of all in-situ/conventional data, passive microwave (MW) radiances, infrared (IR) radiances, and Global Navigation Satellite Systems (GNSS) radio occultation (RO) data compared to a Control that uses the full observing system. The respective experiments use the same analyses for both forecasting systems, produced with the physics-based IFS.

Results highlight that both systems show significant forecast degradation when the selected observing systems are denied, with degradations over the troposphere that are mostly ball-park similar to degradations previously demonstrated in physics-based systems only. The denial of conventional data leads to the largest degradations over the Northern Hemisphere in both systems, and the denial of MW radiances is most detrimental over the Southern Hemisphere. However, the degradations in the AIFS tend to be larger than the ones found in the IFS, at least for the denial of the conventional data, MW or IR radiances. The largest increase in the degradation in the AIFS is found in the tropical troposphere, where impacts in terms of percentage increase of the forecast error can be doubled compared to the IFS. This is also the area for which the AIFS Control shows the largest benefits compared to the IFS Control. Some of the largest increases in impact in the AIFS are found for the conventional data, tentatively linked to larger changes in the distribution of the input data. The larger vulnerability to the loss of observations in the AIFS is, however, smaller than the benefit in forecast scores that the AIFS brings over the IFS for the levels and parameters considered. GNSS-RO data show a very strong forecast impact in the upper troposphere/lower stratosphere in the IFS, particularly in the tropics, and this is more muted in the AIFS, likely due to the focus on the troposphere during the training of AIFS.

Plain Language Summary

This study looks at what happens if observations are withheld from two different weather forecast systems, one using a physics-based description of the atmosphere and one derived from past weather analyses using machine-learning methods. Both systems rely on a gridded representation of the atmosphere as initial conditions, and for the systems considered here these are produced using a traditional physics-based weather forecasting system. The study finds that forecasts in both systems get worse when observations are withheld, with the denial of in-situ observations (weather stations and balloons, aircraft data, etc) leading to the largest degradation over the Northern Hemisphere extra-tropics, whereas the denial of microwave radiances from satellites gives the largest degradation over the Southern Hemisphere. The forecast degradations from the denial of observations tend to be larger for the machine-learning based system, at least for most of the observations considered. However, the machine-learning based forecasts are overall more skillful than their physics-based counterparts, and the larger degradations are still outweighed by this better performance.

1 Introduction

Recent years have seen the advent of data-driven weather forecast models, fueled by the rise of machine learning applications (e.g., Pathak *et al.*, 2022; Keisler, 2022; Lam *et al.*, 2023; Chen *et al.*, 2023; Bi *et al.*, 2023; Nguyen *et al.*, 2023; Lang *et al.*, 2024). Trained on historical data, most commonly the ERA5 reanalysis (Hersbach *et al.*, 2020), these models have shown the potential to outperform leading physics-based global weather prediction systems over many standard forecast scores (e.g., Bouallègue *et al.*, 2024). In addition, forecasts can be produced more quickly (within minutes on a single Graphics Processing Unit - GPU), making such models highly attractive for operational applications. Since 25 February 2025, ECMWF is producing operational weather forecasts with their own data-driven model, known as the Artificial Intelligence Forecasting System (AIFS, Lang *et al.*, 2024). Like all of the data-driven models mentioned here, AIFS currently relies on the physics-based Integrated Forecasting System (IFS, e.g., Rabier *et al.*, 2000) to generate the initial conditions, and the quality and characteristic of these initial conditions crucially depends on the observations that have been assimilated in the IFS.

The present paper evaluates the forecast impact of a range of leading observing systems in IFS as well as AIFS. Using a set of observing system experiments (OSEs) in which observations are denied from a Control with the full observing system, we compare the impacts in the two systems to highlight similarities and differences. As the global observing system is constantly changing, results from such a comparison are important to identify operational vulnerabilities, and in particular to establish whether these are larger or smaller in the data-driven system compared to the more established physics-based systems. To our knowledge, it is the first time that such OSEs are performed systematically across several observing systems.

Observing system experiments with physics-based systems are performed regularly at different Numerical Weather Prediction (NWP) centres, highlighting the importance of a range of observations for forecast quality (e.g., Bouttier and Kelly, 2001; Lawrence *et al.*, 2019; Bormann *et al.*, 2019; WMO, 2024). Results of these are reviewed at dedicated impact workshops organised every four years by WMO, providing important input for the evolution of the global observing system (e.g., WMO, 2024). While there are some variations in impact reported by different NWP centres, for instance due to the number of data used or the sophistication of the assimilation methods, the leading observing systems for global weather forecasts are typically the in-situ data (including radiosondes, synop stations, aircraft data, etc), satellite radiances from passive microwave (MW) or infrared (IR) sensors, as well as radio occultation (RO) observations based on Global Navigation Satellite Systems (GNSS). Other observations such as Atmospheric Motion Vectors (AMVs) or scatterometer data are also playing important roles, not least for tropical cyclone forecasting.

Forecasts with machine-learning based forecast models may show different sensitivity to the loss of observations than physics-based systems for two main reasons. Firstly, the reduction in model error achieved with these systems may mean that forecasts are more sensitive to increased errors in the initial conditions. Secondly, data-driven systems rely on the characteristics of the input data used for inference to be similar to those of the training data, and shifts in the distribution of the input data can lead to unexpectedly poor performance. Such a situation was encountered, for instance, during pre-operational testing of the move from cycle 49r1 to 50r1 at ECMWF. AIFS initially showed significantly poorer performance with analyses produced with cycle 50r1 than with 49r1, not reflecting that the analysis produced with cycle 50r1 was more accurate. Fine-tuning of the AIFS with 50r1 analyses was needed to adapt to the changed characteristics of the analyses. As the denial of entire observing systems constitutes a similar perturbation of the analysis characteristics, it is hence an open question how AIFS reacts to such a change, without performing fine-tuning.

The structure of the paper is as follows: we first introduce the two forecasting systems and experiments used in the present study. We then contrast the forecast performance of the two systems when both benefit from the full set of observations, before comparing the impact of denying observations in each system. The impact results are complemented by a characterisation of how mean analyses change as a result of denying different observations. Conclusions are presented in the last section.

2 Forecasting systems and experiments

2.1 IFS

IFS-based analyses and forecasts investigated in this paper were generated with cycle 49r1, that is, the cycle that was used operationally at ECMWF between 12 November 2024 and 11 May 2026. The experiments cover the period 1 June – 30 August 2022 and 1 December 2022 – 28 February 2023, and analyses were produced with ECMWF's 12-hour 4-dimensional Variational (4D-Var) analysis system, configured with a spatial model resolution of T_{CO} 399 ($\approx$ 27 km), a final incremental analysis resolution of T_L 255 ($\approx$ 80 km), and 137 levels in the vertical. Flow-dependent background errors are specified through an Ensemble of Data Assimilations (EDA, e.g., Bonavita *et al.*, 2012). Changes in the size of the background errors due to denying observations are neglected, as previous work showed only minor sensitivity to this for the kind of denials that are considered (Duncan *et al.*, 2021). The spatial model resolution used here is coarser than the 9 km resolution used operationally, to save computational cost. Ten-day forecasts are produced every 12-hours.

The following assimilation experiments were run:

Control: Full observing system as used operationally at the time. This encompasses a wide range of in-situ and satellite data, as summarised in Table 1.

No conventional: As Control, but with all in-situ data denied. This includes all non-satellite data apart from ground-based radar, such as radiosondes, synop stations, aircraft data, etc. For traditional reasons, we will refer to these as ”conventional” observations, even though we recognise that satellite data are by now also quite conventional.

No MW radiances: As Control, but excluding all satellite radiances from passive MW sensors (imagers and sounders).

No IR radiances: As Control, but excluding all satellite radiances from passive IR sensors (hyperspectral IR sounders as well as geostationary imagers).

No GNSS-RO: As Control, but excluding all GNSS-RO data.

2.2 AIFS

The AIFS forecasts for this study came from the operational version 1.1.0 (active from 27 August 2025 to 12 May 2026) of the AIFS-Single described in Moldovan *et al.* (2026). This version of the model uses an encoder-processor-decoder architecture, using a graph-based attention encoder/decoder and a transformer processor with sliding window attention (Lang *et al.*, 2024). The model operates on a reduced Gaussian grid at the N320 resolution ($\approx$ 31 km), the upper atmosphere is represented by geopotential, horizontal wind components, specific humidity, and temperature at 13 pressure levels: 50, 100, 150, 200, 250, 300, 400, 500, 600, 700, 850, 925, and 1000 hPa. During training, a pressure-level weighting is applied which decreases linearly with height, so that upper atmospheric levels contribute less to the total loss. The forecasts are initialised from the different IFS experiments outlined in Section 2.1 and run with a temporal resolution of 6-hours and a lead time of 10 days.

Table 1: Observing systems used in the Control experiment.

Observing system	Satellites, instruments or observation types
Conventional	Synops, ship reports, aircraft data, drifting buoys, radiosondes, pilot balloons and profilers
Microwave radiances	AMSU-A from NOAA-15, -18, -19; Metop-B, -C ATMS from S-NPP, NOAA-20 MHS from NOAA-19; Metop-B, -C MWHS-2 from FY-3C, D, E SSMIS from DMSP-17, -18 AMSR2 from GCOM-W1 GMI from GPM
Infrared radiances	AIRS from Aqua IASI from Metop-B, -C CrIS from S-NPP, NOAA-20 Geostationary All-sky Radiances (ASRs) from Meteosat-9, -11; Himawari-8, -9; GOES-16, -17, -18
GNSS-RO bending angles	From Cosmic-2, GRACE-2, KOMPSAT-5, Metop-B, -C; Sentinel-6a, Spire, TANDEM-X, TerraSarX
AMVs	From Meteosat-9, -11; Himawari-8, -9; GOES-16, -17, -18; NOAA-15, -18, -19; Metop-B, -C; S-NPP, NOAA-20, Aqua
Scatterometer	From Metop-B, -C; HY-2B
Other	Horizontal line of sight wind profiles from Aeolus Rain rates from ground-based radar Ozone retrievals from GOME-2, SBUV, OMI

2.3 Verification

All experiments are verified against ECMWF's operational analyses, and a range of geophysical parameters, regions and forecast ranges have been considered. The main focus is the troposphere, primarily given the limited representation of the stratosphere in the AIFS. The operational analysis is chosen here as reference as it represents the best estimate of the atmospheric state, produced by a system that ran independently of the experiments performed here. Note, however, that analysis-based forecast verification can lead to mis-leading results for short forecast ranges, when errors in the analyses and forecasts may not be independent (e.g. Bormann *et al.*, 2019). For this reason, we focus on forecast ranges of day 2 and beyond. To allow the analyses to adjust to the denial of the respective observation types, we exclude the first 4 days of the experiment periods in the statistics presented.

3 Results

3.1 Comparison of IFS and AIFS for the Control experiment

Before we evaluate the impact of the observing systems considered here, it is worth contrasting the forecast performance of the Control experiments for the two systems considered, in order to provide context for the assessment that follows.

For the troposphere, the AIFS clearly outperforms the IFS for key geophysical parameters on the pressure

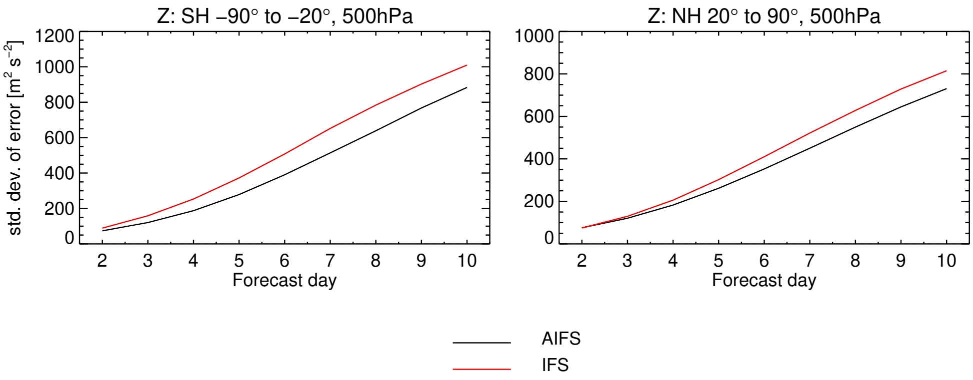


Figure 1: Standard deviation of the forecast error of the 500 hPa geopotential height as a function of forecast range for the Control from the AIFS (black) and IFS (red), over the Southern (left) and Northern (right) Hemisphere extra-Tropics. Verification is against the operational analysis and results from both experiment periods are combined.

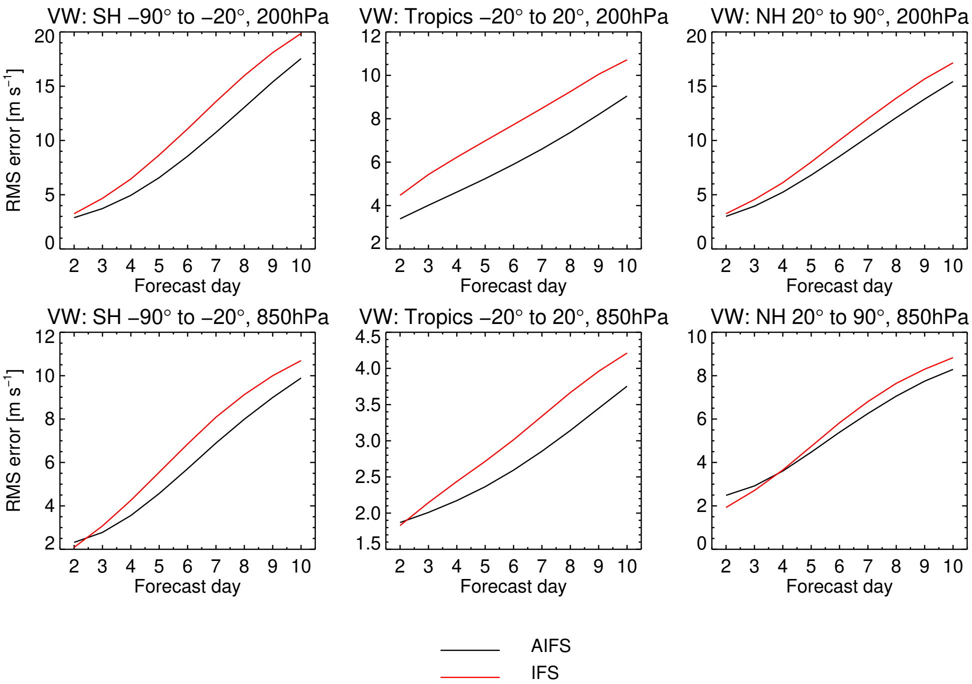


Figure 2: Root means square of the wind forecast error at 200 hPa (top row) and 850 hPa (bottom row) as a function of forecast range for the Control from the AIFS (black) and IFS (red), over the Southern Hemisphere extra-Tropics (left), the Tropics (middle) and the Northern Hemisphere extra-Tropics (right). Verification is against the operational analysis and results from both experiment periods are combined.

levels covered by AIFS, as highlighted in Figures 1 and 2. The reduction in forecast error is mostly between 10 and 25 % compared to the IFS at mid and upper tropospheric levels for the parameters shown, translating to a gain in forecast skill over the IFS of around half a day for the Northern Hemisphere extra-Tropics at day 5, and about one and $1\frac{1}{2}$ days for the Southern Hemisphere and the Tropics, respectively. Substantially better performance of the AIFS system is also seen in a range of other forecast metrics, and overall they are not achieved at the expense of smoother forecasts.

For the stratosphere, the performance of the AIFS is less convincing, for instance, exhibiting mostly larger forecast errors than the IFS at 50 hPa (not shown). This is a recognised feature of the AIFS version used, partly the result of linear loss scaling with altitude during the training, hence placing less emphasis on forecast performance for these levels (Lang *et al.*, 2024).

The results shown here are in line with previous findings from Lang *et al.* (2024), reflecting significant benefits in forecast skill for the AIFS system.

3.2 Overall impact of different observing systems in the IFS and AIFS

We will now evaluate the impact of the different observing systems by assessing the change in forecast error resulting from the denial of observations relative to the Control. This will be done for the IFS and the AIFS forecasts, respectively, and the results contrasted against each other.

For the troposphere and the extra-Tropics, the impact of the four observing systems in both forecasting systems is broadly in line with that reported for physics-based systems in previous studies, and the forecast impacts are of similar orders of magnitude. Considering the 500 hPa geopotential (Fig. 3), the conventional observations show the largest impact over the Northern Hemisphere extra-Tropics, the

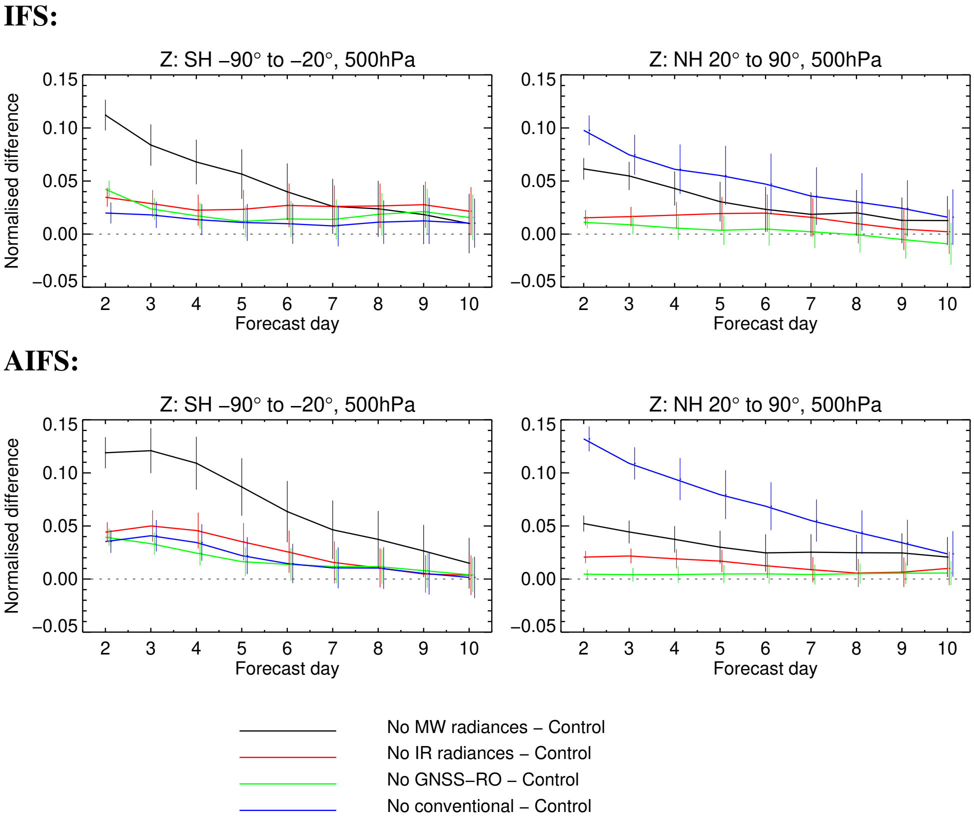


Figure 3: Normalised difference in the standard deviation of the 500 hPa forecast error for the various denial experiments compared to the Control as a function of forecast range over the Southern (left) and Northern (right) Hemisphere extra-Tropics. Verification is against the operational analysis and results for the two periods considered are combined. Vertical bars indicate statistical significance at the 95 % confidence level. Different colours indicate different experiments, as shown in the legend. The top row shows results from the IFS, whereas the bottom row shows results for the AIFS.

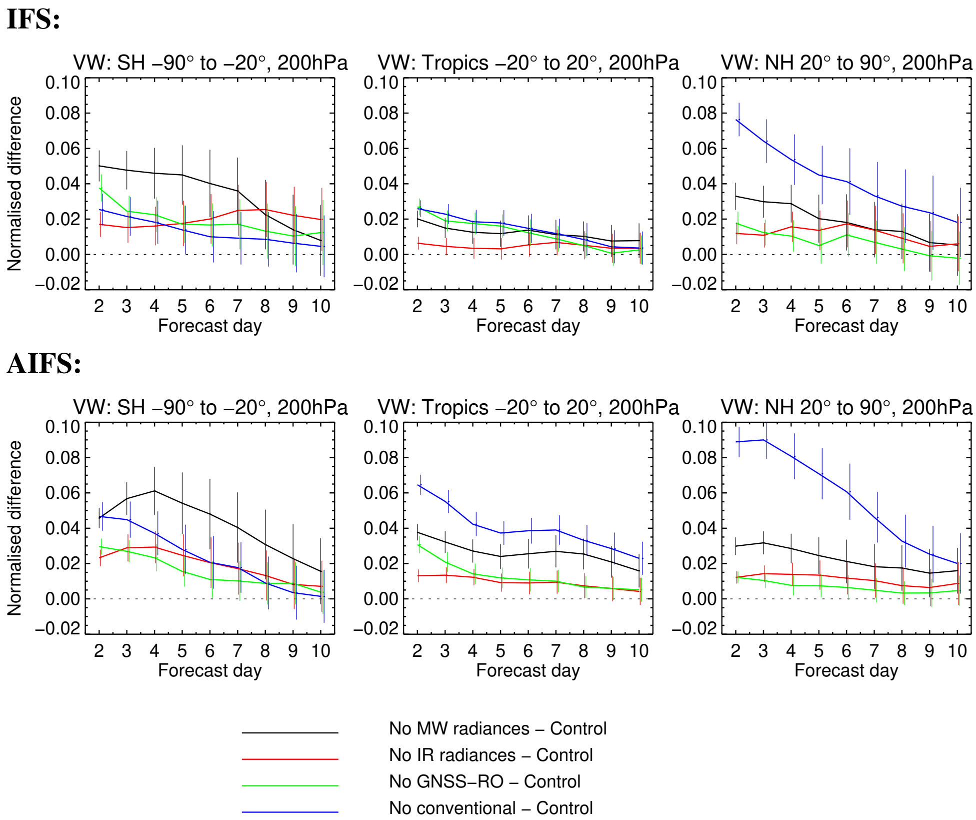
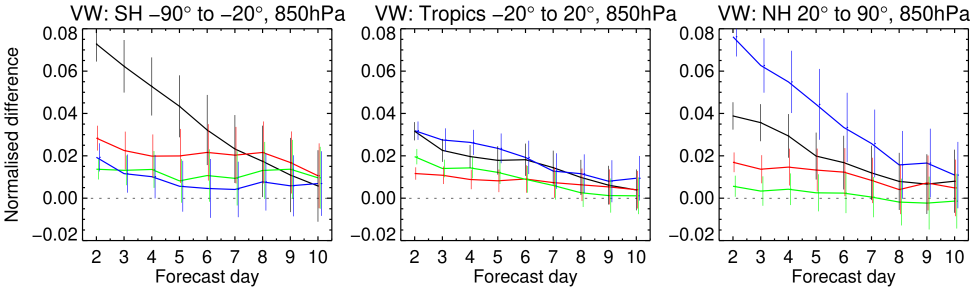


Figure 4: Normalised difference in the root mean square error for the 200 hPa wind forecast for the various denial experiments compared to the Control as a function of forecast range over the Southern Hemisphere extra-Tropics (left), the Tropics (middle) and the Northern Hemisphere extra-Tropics (right). Verification is against the operational analysis and results for the two periods considered are combined. Vertical bars indicate statistical significance at the 95 % confidence level. Different colours indicate different experiments, as shown in the legend. The top row shows results from the IFS, whereas the bottom row shows results for the AIFS.

region best-covered by such observations. MW radiances show the largest impact over the Southern Hemisphere extra-topics and generally provide the largest benefit of the satellite data considered here. IR radiances and GNSS-RO data also provide statistically significant benefits for the 500 hPa geopotential, particularly over the Southern Hemisphere extra-Tropics. The findings are replicated for other geophysical variables and levels in the troposphere, as seen, for instance, for wind forecast errors at 200 and 850 hPa (Figures 4 and 5), with some variations, reflecting the different strengths of the observing systems considered. Compared to OSEs performed seven years ago (Bormann *et al.*, 2019), results for the IFS are mostly remarkably similar, with the main differences being a larger impact from GNSS-RO data in the present study, expected due to the increase of the number of observations assimilated, and a slight reduction in the impact of microwave radiances. However, a strict comparison of results obtained over different time periods can be misleading, due to seasonal and inter-annual fluctuations in impact, so these comparisons should be treated with some caution.

Comparing tropospheric impacts in the IFS and the AIFS more closely, there is nevertheless a tendency for larger impact in the AIFS, in terms of percentage of error increase when the data are denied, and this is most notable in the Tropics. This can be more clearly seen in Fig. 6 which highlights that the size of the day 5 forecast impact of conventional observations as well as MW and IR radiances on wind at 200 hPa

IFS:



AIFS:

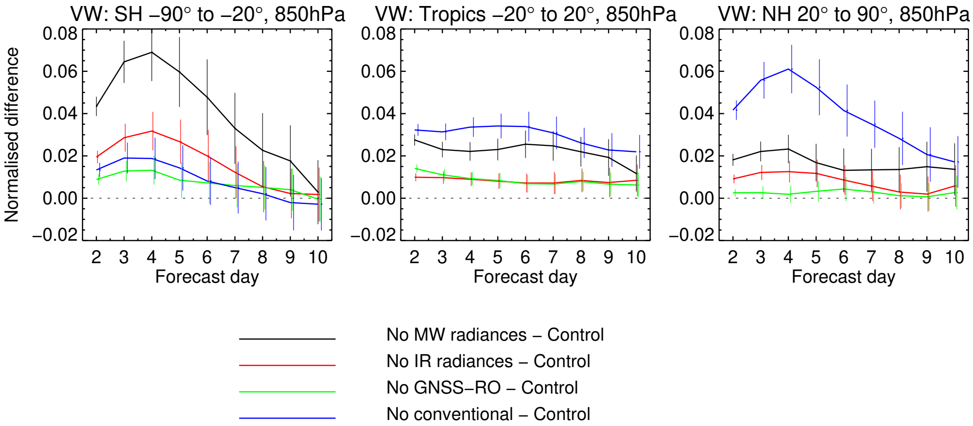


Figure 5: As Fig. 4, but for the 850 hPa wind forecast.

over the Tropics is at least twice that of the impact in the IFS. The impact is also lasting longer into the forecast, staying statistically significant over the full forecast range for the conventional observations and MW radiances (Fig. 4). For the extra-Tropics, there is also a tendency for larger and longer-lasting impact in the AIFS for conventional observations, MW and IR radiances, particularly for the Southern Hemisphere, but the differences are not as marked as over the Tropics (see, for instance, Figures 6 and 7). The most consistent and largest increase in impact in the AIFS over the regions considered is seen for conventional observations, and in the AIFS they are the most impactful data not only over the Northern Hemisphere, but also the Tropics.

It is notable that the larger impact in the AIFS occurs over the regions for which the AIFS Control shows the largest advantage over the IFS Control in terms of forecast performance, in particular over the Tropics. One interpretation is that the the growth of errors in the initial conditions is becoming more relevant in the AIFS, as error growth due to model errors is reduced. Hence larger errors in the initial conditions resulting from denying observations have a relatively larger effect on forecast errors. Other factors may also play a role, such as changes to the mean or standard deviation of the analyses, which can have a more significant impact in data-driven systems, and these will be explored in section 3.4. Note, however, that the largest degradations noted here, either due to the loss of MW radiances over the Southern Hemisphere or the loss of conventional data over the Northern Hemisphere, both translate to a loss of forecast skill in AIFS of around 6 h at day 5. The degradations seen are hence relatively small compared to the clear advantages that the AIFS has compared to the IFS in terms of tropospheric forecast scores. So the slightly larger vulnerability to changes in the observing system in AIFS is still outweighed by the overall better forecast performance.

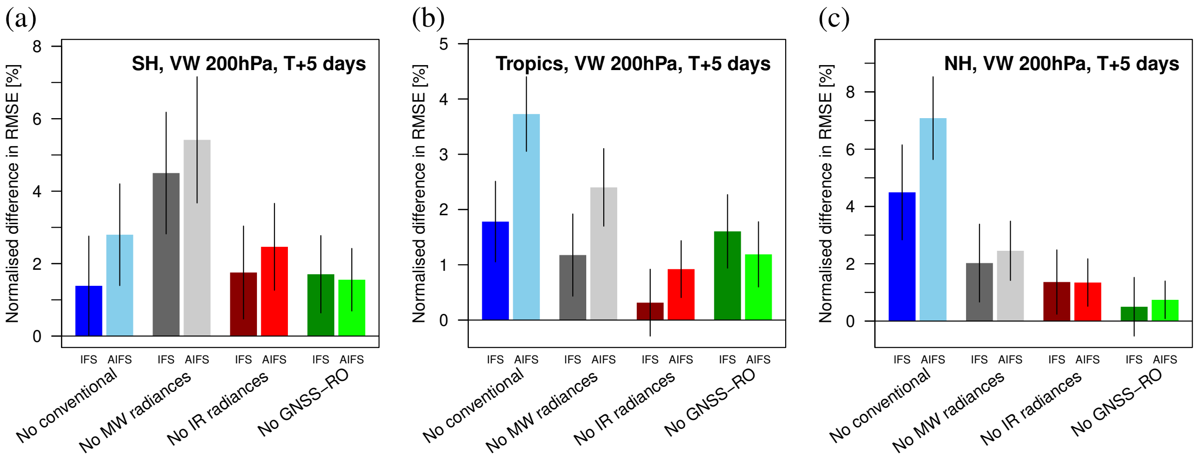


Figure 6: Normalised difference in the root mean square error for the day 5 200 hPa wind forecast for the various denial experiments for IFS (darker colours) and AIFS (lighter colours) compared to the respective Control over a) the Southern Hemisphere extra-Tropics, b) the Tropics (middle) and c) the Northern Hemisphere extra-Tropics. Verification is against the operational analysis and results for the two periods considered are combined. Vertical bars indicate statistical significance at the 95 % confidence level. Different colours indicate the different denial experiments, as indicated on the x-axis.

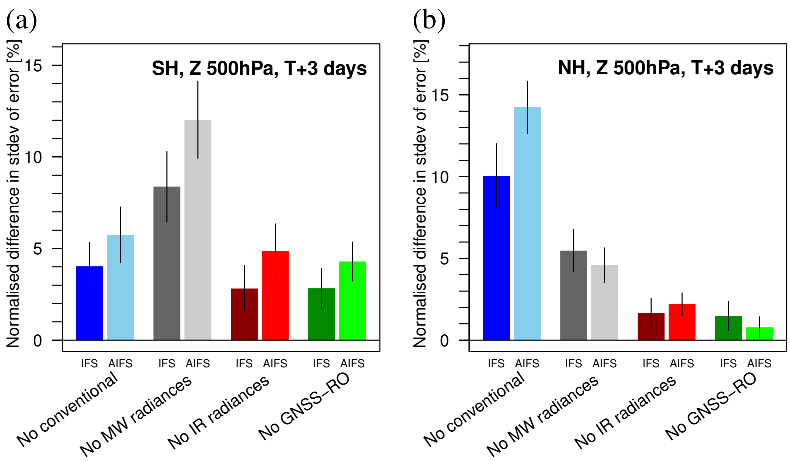


Figure 7: Similar to Fig. 6, but for the standard deviation of the day 3 forecast error for the 500 hPa geopotential over a) the Southern Hemisphere extra-Tropics, and b) the Northern Hemisphere extra-Tropics.

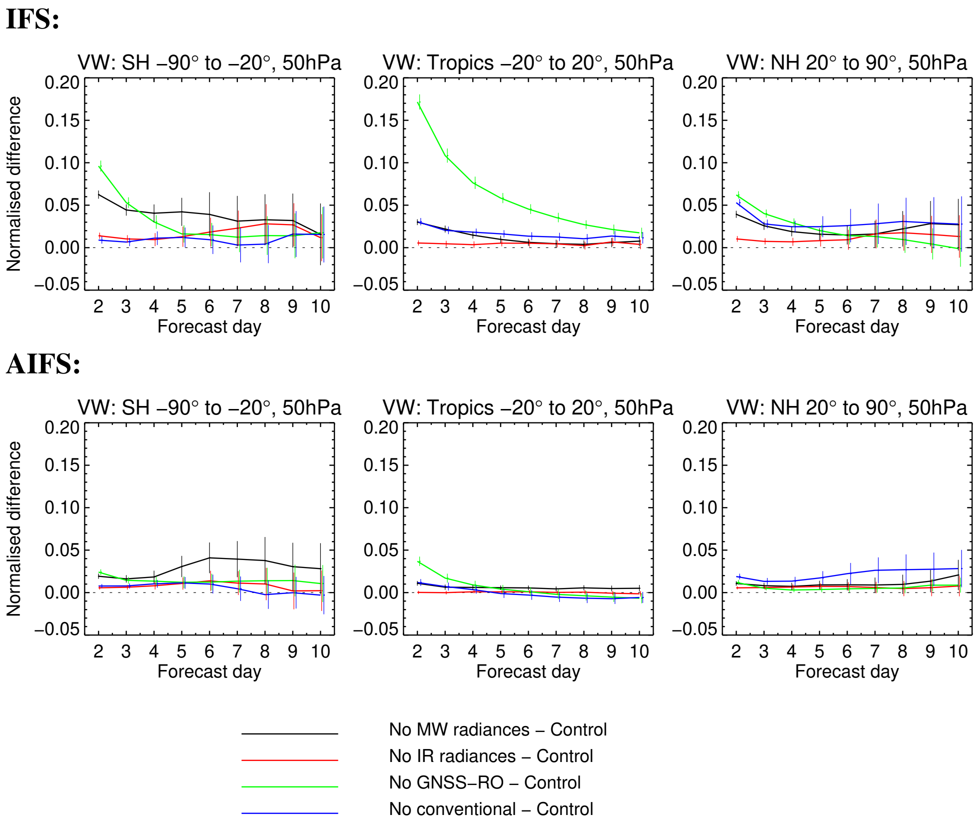


Figure 8: As Fig. 4, but for the 50 hPa wind forecast.

For GNSS-RO data, the tropospheric impact observed in the AIFS and the IFS is overall relatively similar, in contrast to the changes seen for the other observing systems. While GNSS-RO data provides leading impacts in the upper troposphere in the IFS, for instance for wind (Fig. 4) or temperature (not shown), GNSS-RO does not experience the same increase in impact in the AIFS as the other observing systems. GNSS-RO observations have the largest sensitivity in the upper troposphere/lower stratosphere and add important temperature information with high vertical resolution as well as anchoring for radiance bias corrections in the IFS (Bauer *et al.*, 2014). It is possible that some of these aspects play a smaller role in the AIFS, for instance, due to the emphasis on lower levels during the training of AIFS or the very limited representation of stratospheric levels in the AIFS. The finding should be reevaluated with future versions of the AIFS with more stratospheric levels.

The limited representation of the stratosphere in the AIFS is likely also be the reason why the forecast impact of the observing systems considered is very different between the AIFS and IFS over the stratosphere. This can be seen, for instance, in Fig. 8 for wind at 50 hPa - the only level above 100 hPa included in this version of the AIFS. For the IFS, the large forecast impact of GNSS-RO data over the Tropics stands out here, a feature that has been noted in previous OSEs (Bormann *et al.*, 2021) and has been becoming more prominent as a result of the growing number of GNSS-RO observations used (e.g., Anthes *et al.*, 2024). While GNSS-RO also shows the largest impact at day 2 and 3 in the AIFS in this region, the effect is much more muted and mostly lost beyond day 4. Similarly, there is relatively little resemblence in the forecast impact for the other observing systems considered here, including over the extra-Tropics. The finding is likely linked to the poorer performance of the AIFS compared to the IFS at 50 hPa in general.

3.3 Seasonal aspects

There are some weak seasonal differences in the impacts of the observing systems considered here in both systems (Fig. 9). For the IFS, the conventional data tends to have the leading impact over the Northern Hemisphere in winter, whereas during summer microwave radiances and conventional data show more comparable impacts. This is consistent with previous findings reported in Bormann *et al.* (2019) and Lawrence *et al.* (2019), and some of the weaker impact of the microwave data has been attributed to a poorer use of the data over sea-ice regions. The seasonal behaviour is, however, not replicated over the Southern Hemisphere, for which the impact of the MW data for the two seasons is comparable or slightly larger during the Southern Hemisphere winter.

These seasonal differences are only partially replicated in the AIFS system. Here, the impacts of all observing systems are notably larger during the June–August period over both hemispheres. In addition, the conventional data are the leading observing system over the Northern Hemisphere also during the summer period, for which the gain in impact in the AIFS system is particularly striking. The reasons for these slightly different behaviours are not fully understood. It is notable that AIFS performs particularly well compared to the IFS over the June–August period for both hemispheres. For this period, the percent reduction in forecast errors for the AIFS compared to IFS is typically twice as large as for the December–February period (not shown). So a larger reduction in model errors for the June–August period may contribute to a larger sensitivity to errors in the initial conditions. However, seasonal differences in the changes to the analysis characteristics introduced by denying the four observing systems considered may also play a role.

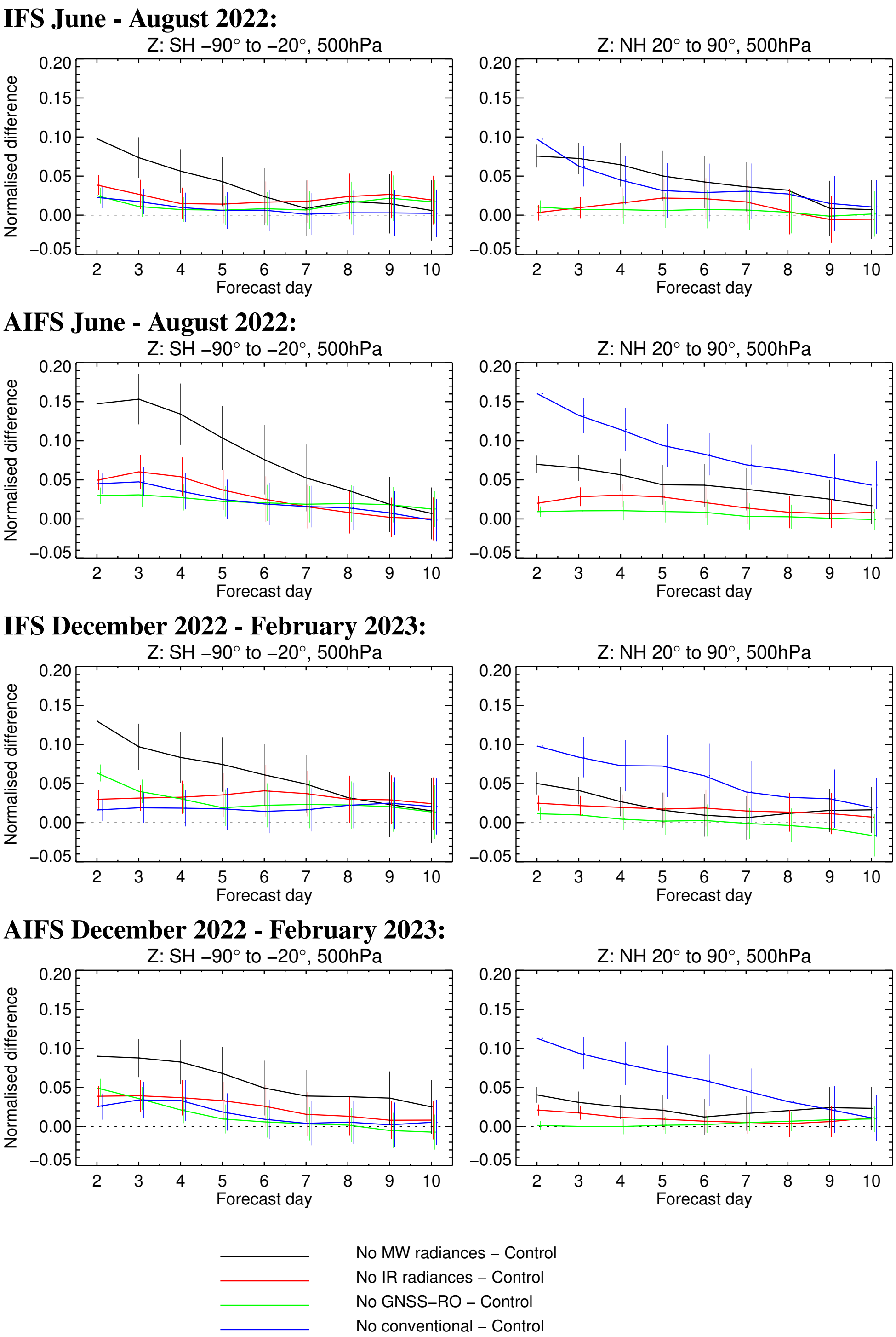


Figure 9: As Fig. 1, but for the June - August 2022 period in the top two rows and the December 2022 - February 2023 period in the bottom two rows.

3.4 Analysis changes

Practical experience with data-driven models has shown that inference results can be sensitive to changes in the distribution of the input data, that is, forecasts can degrade considerably if the initial conditions used for predictions are, in some way, different from those encountered during training. To address this, fine-tuning is often employed, a method that allows to adjust the trained weights to the new data characteristics, while still retaining most of the learned relationships (Lang *et al.*, 2024). No fine-tuning to the altered characteristics in the denials has been performed in the AIFS results presented in this paper. Previous OSEs have shown that the denial of observation types in the IFS can have significant influence on the means of the resulting analyses (Bormann *et al.*, 2019), and this aspect may contribute to the different performance seen in the AIFS. To explore this aspect further, we will now investigate differences in the means of IFS analyses from the OSEs presented here.

Examples of changes in the mean analyses are shown in Figures 10 and 11, for the mean sea-level pressure (MSLP) and temperature. It is striking how the denial of the conventional data leads to the largest changes in the mean analyses for MSLP, with a reduction of up to 2 hPa at high latitudes, and also some of the largest changes for temperature in the troposphere. While all observing systems have some effect on the zonal mean temperature analyses in the troposphere, these are typically below 0.2 K for the radiances or GNSS-RO data, whereas the denial of the conventional data leads to wide-spread cooling for the lower troposphere over the Tropics, exceeding 0.4 K in places, and wide-spread warming in the upper troposphere in the extra-Tropics. There are similar findings for zonal means of the zonal wind component, for which the denial of conventional data leads to significant changes for the tropical mid-troposphere, whereas the denial of the other observing systems has comparatively little effect on the zonal mean analyses (not shown). Some larger changes in the zonal mean temperatures above 200 hPa are seen as a result of denying the GNSS-RO observations (Fig. 11c), with cooling and warming reaching

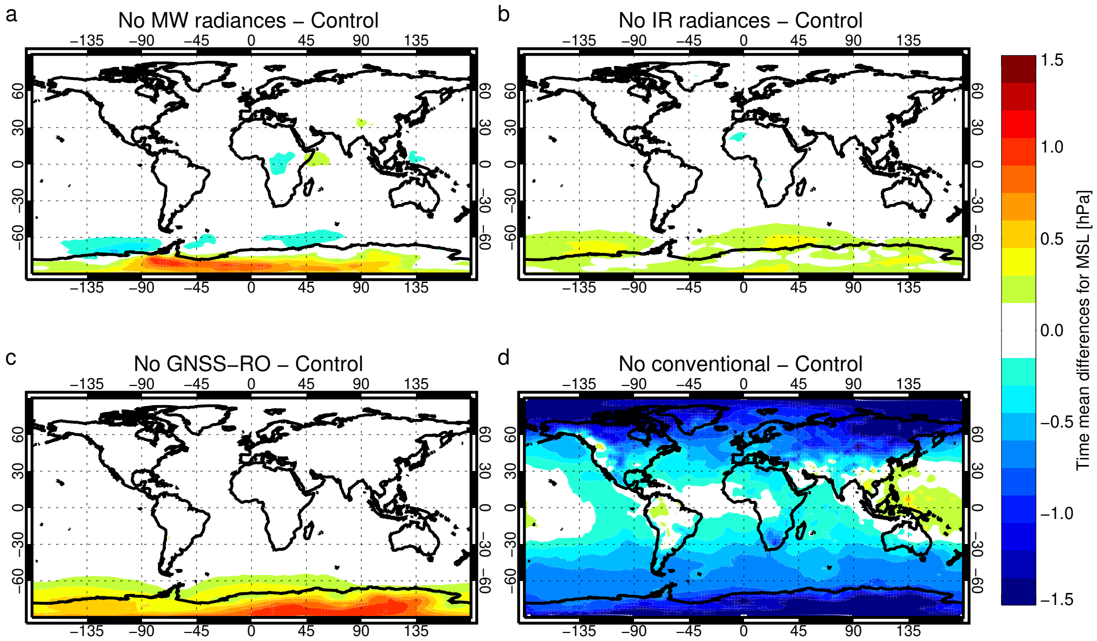


Figure 10: Maps of differences in the average analyses of mean sea level pressure [hPa] compared to the Control for a) the No MW radiances experiment, b) the No IR radiances experiment, c) the No GNSS-RO experiment, and d) the No conventional observations experiment. Results for the two periods are combined.

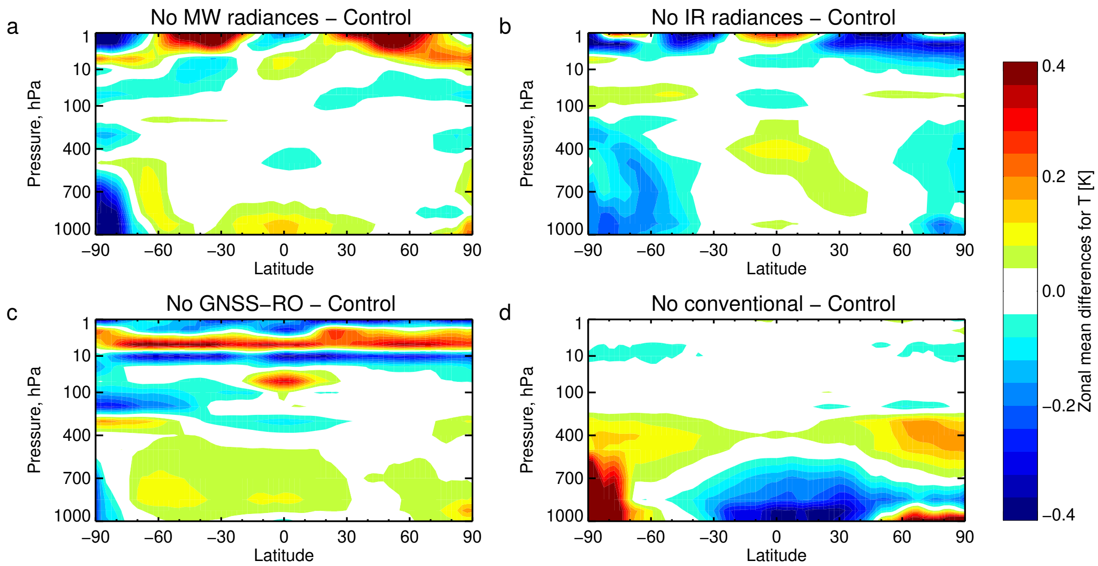


Figure 11: Zonal means differences in the temperature analyses [K] compared to the Control for a) the No MW radiances experiment, b) the No IR radiances experiment, c) the No GNSS-RO experiment, and d) the No conventional observations experiment. Results for the two periods are combined.

0.4 K, but these are mostly confined to levels of 10 hPa and above, and hence not used as input to AIFS. For relative humidity, the largest changes in the zonal mean are seen when the MW radiances are denied, with wide-spread drying approaching 1 % relative humidity over much of the lower and mid troposphere (noth shown).

The finding that changes to the mean analyses for MSLP, temperature and wind are largest for the conventional data provides an additional possible explanation why some of the increases in forecast degradation with the AIFS are largest for conventional observations. It appears that the distribution of the tropospheric analyses is changed the most for this observation type, likely contributing to a poorer performance of AIFS. In this context it is also worth mentioning that the standard deviation of the temperature and wind analyses is also significantly reduced over much of the troposphere in the analyses without conventional data, whereas such prevalent reduction is not seen for the other observing systems (not shown). This further highlights significant differences in the characteristics of the initial conditions used in the case of the conventional data and may hence contribute to a poorer performance. At the same time, differences in the mean or standard deviation of the analyses are present over both seasons considered (not shown), whereas the June–August period shows stronger degradations in the AIFS than over the December–February period. So it is clear that other factors are also affecting the size of the larger forecast degradation seen in the AIFS from the denial of observations, such as the reduction in model error achieved in the AIFS compared to the IFS.

4 Conclusions

This paper summarised first experiences with OSEs for the data-driven AIFS and intercompared these to findings with the physics-based IFS. The analyses used to generate the forecasts in each of the denial experiments are the same for both systems, and they are produced with the physics-based 4D-Var of the IFS by denying selected observations from a system that otherwise uses the full observation complement.

For AIFS, the same model is used for all AIFS forecasts and no fine-tuning to the analyses with a reduced observation set has been performed. The observing systems considered are conventional observations, satellite data from passive MW sensors, IR radiances, and bending angles from GNSS-RO. The main findings are:

- The tropospheric forecast impact from denying the observing systems considered is ball-park similar for the two forecasting systems considered, though the denial of observations tends to lead to a larger forecast degradation in the AIFS than in the IFS for conventional observations, MW and IR radiances.
- The largest increase in degradation in the AIFS is found for the tropical troposphere, where impacts can be doubled for the conventional data, MW or IR radiances compared to the IFS. The Tropics are also the area where the gain in forecast skill when moving from IFS to AIFS is largest.
- The most consistent and largest increase in impact in the AIFS is found for conventional observations. This may be linked to larger changes in the mean and standard deviation of the analysis fields compared to the Control for this denial, and these changes in the characteristics of the initial conditions may contribute to a poorer performance.
- In contrast to the findings for other observing systems, the impact of GNSS-RO data in the troposphere is similar in the AIFS and IFS forecasts. At the same time, the large impact of GNSS-RO in the lower stratosphere in the IFS is severely diminished with AIFS. The latter likely reflects a poor representation of the stratosphere in AIFS due to the focus on the troposphere during training, and diminished impacts in the stratosphere are also seen for the other observing systems considered.
- For the IFS, results from the current OSEs are largely similar to those obtained seven years ago by Bormann *et al.* (2019), with the most notable difference being an increase in the impact of GNSS-RO data, likely due to significantly increased observation numbers.

The slightly larger vulnerability of the tropospheric AIFS forecasts to the denial of observations is likely due to the combination of two factors: 1) a larger influence of initial-condition errors due to reduced model error, and 2) changes in the distribution characteristics of the analyses having a negative effect on forecast performance for the AIFS. The latter aspect may well dominate, particularly for the denials with the largest changes to the analysis characteristics. To investigate this point further, it would be interesting to perform fine-tuning with analyses from the respective denials, to ascertain to what extent this allows recuperating some of the lost forecast performance. However, in any case, the larger vulnerability of the AIFS to changes in the observing system is not dramatic, particularly considering that the denials investigated here represent drastic changes, and it is outweighed by the better performance of AIFS compared to the IFS for the forecast scores considered. This is important for operational application of AIFS on its own. It is also reassuring for potential future operational applications of combining AIFS and IFS via spectral nudging: it is unlikely that typical changes to the observing system, such as the loss of a single satellite or sensor, will introduce a performance degradation in the nudged IFS that defeats the benefit otherwise introduced through the nudging to AIFS at larger scales.

While the performance of AIFS in the OSEs considered here is reassuring for operational applications, a similar level of robustness in OSEs still needs to be established for the AI-DOP (Direct Observation Prediction) system currently under development at ECMWF (e.g. McNally *et al.*, 2024; Alexe *et al.*, 2024). In the AI-DOP approach, the forecast model is directly trained from observations alone, without the use of a physics-based system. Initial OSEs with a very early version of AI-DOP showed much larger degradations in forecast skill compared to the IFS when selected observations were denied at inference stage, up to two orders of magnitude larger than encountered in the IFS (Bormann *et al.*, 2025). It was hypothesised that this larger vulnerability can be reduced either through a revised training regime that more systematically accounts for potential observation drop-outs, or through fine-tuning to reflect the altered observing system used at inference. Further work is needed to confirm that these or other modifications can indeed address the large vulnerability to observation losses in AI-DOP.

The OSEs presented in this paper serve as a reminder that observations are crucial for making accurate weather forecasts, whether they are generated with a machine-learning or a physics-based forecast model. It remains to be seen to what extent machine-learning applications will change the relevance and impact of certain observation types, an important aspect for the future evolution of the observing system. In this respect, the current findings suggest that experiences with physics-based NWP systems also provide guidance for machine-learning based systems, at least as long as they are based on analyses generated with physics-based system. However, developments such as AI-DOP have the potential to alter this, as they may allow the use of observations that have so far been elusive for physics-based systems, for instance due to complexities in the physical modelling of underlying processes. Similarly, machine learning approaches integrated into physics-based systems also have the potential to allow a more complete exploitation of observations that are currently too difficult to handle (e.g., Geer, 2024), and in the future this may also influence the observation impact that can be achieved with a given observing system. A comprehensive observing system that evolves with and caters for these different needs continues to be essential for NWP.

Acknowledgements

IFS and AIFS are systems developed by a large number of past and present contributors at ECMWF and beyond, and their contributions are gratefully acknowledged.

References

Alexe, M., Boucher, E., Lean, P., Pinnington, E., Laloyaux, P., McNally, A., Lang, S., Chantry, M., Burrows, C., Chrust, M., Pinault, F., Villeneuve, E., Bormann, N. and Healy, S. (2024). GraphDOP: Towards skilful data-driven medium-range weather forecasts learnt and initialised directly from observations. *arXiv preprint arXiv:2412.15687*, doi:10.48550/arXiv.2412.15687, URL https://arxiv.org/abs/2412.15687.

Anthes, R. A., Marquardt, C., Ruston, B. and Shao, H. (2024). Radio occultation modeling experiment (romex): Determining the impact of radio occultation observations on numerical weather prediction. *Bulletin of the American Meteorological Society*, **105**(8), E1552 – E1568, doi:10.1175/BAMS-D-23-0326.1, URL https://journals.ametsoc.org/view/journals/bams/105/8/BAMS-D-23-0326.1.xml.

Bauer, P., Radnóti, G., Healy, S. and Cardinali, C. (2014). GNSS radio occultation constellation observing system experiments. *Monthly Weather Review*, **142**(2), 555–572, doi:10.1175/MWR-D-13-00130.1, URL https://journals.ametsoc.org/view/journals/mwre/142/2/mwr-d-13-00130.1.xml.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X. and Tian, Q. (2023). Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, **619**, 533–538, doi:10.1038/s41586-023-06185-3.

Bonavita, M., Isaksen, L. and Hólm, E. (2012). On the use of EDA background error variances in the ECMWF 4D-Var. *Quart. J. Roy. Meteor. Soc.*, **138**(667), 1540–1559, doi:10.1002/qj.1899, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.1899.

Bormann, N., Burrows, C., Day, J., Duncan, D., English, S., Geer, A., Healy, S., Ingleby, B., Laroche, S., Lawrence, H., Lean, K., Lonitz, K., Lupu, C., McNally, T., Poan, E., Randriamampianina, R., Rennie, M. and Sandu, I. (2021). Evaluating forecast impact of current and future observations. In *ECMWF Annual Seminar 2021: Observations*, URL https://events.ecmwf.int/event/217/contributions/2049/attachments/1397/2509/AS2021_Bormann.pdf.

Bormann, N., Lawrence, H. and Farnan, J. (2019). Global observing system experiments in the ECMWF assimilation system. *Technical Report 839*, ECMWF Tech. Memo., doi:10.21957/sr184iyz, URL https://www.ecmwf.int/node/18859.

Bormann, N., McNally, T., Alexe, M., Lean, P., Boucher, E., Lang, S., Pinnington, E. and Lessig, C. (2025). Skilful weather predictions from observations alone: the role of passive sounders. In *25th International TOVS Study Conference, Goaa, India*, URL https://itwg.ssec.wisc.edu/wordpress/wp-content/uploads/2025/06/oral.7.02.Bormann_itsc25.pdf.

Bouallègue, Z. B., Clare, M. C. A., Magnusson, L., Gascón, E., Maier-Gerber, M., Janoušek, M., Rodwell, M., Pinault, F., Dramsch, J. S., Lang, S. T. K., Raoult, B., Rabier, F., Chevallier, M., Sandu, I., Dueben, P., Chantry, M. and Pappenberger, F. (2024). The rise of data-driven weather forecasting: A first statistical assessment of machine learning–based weather forecasts in an operational-like context. *Bulletin of the American Meteorological Society*, **105**(6), E864 – E883, doi:10.1175/BAMS-D-23-0162.1, URL https://journals.ametsoc.org/view/journals/bams/105/6/BAMS-D-23-0162.1.xml.

Bouttier, F. and Kelly, G. (2001). Observing-system experiments in the ECMWF 4-DVAR assimilation system. *Quart. J. Roy. Meteor. Soc.*, **127**, 1469–1488.

Chen, L., Zhong, X., Zhang, F., Cheng, Y., Xu, Y., Qi, Y., and Li, H. (2023). Fuxi: a cascade machine learning forecasting system for 15-day global weather forecast. *npj Climate and Atmospheric Science*, **6**(1), doi:10.1038/s41612-023-00512-1.

Duncan, D. I., Bormann, N. and Hólm, E. (2021). On the addition of microwave sounders and numerical weather prediction skill. *Quart. J. Roy. Meteor. Soc.*, **147**(740), 3703–3718, doi:10.1002/qj.4149, URL https://doi.org/10.1002/qj.4149.

Geer, A. J. (2024). Simultaneous inference of sea ice state and surface emissivity model using machine learning and data assimilation. *Journal of Advances in Modeling Earth Systems*, **16**(7), e2023MS004080, doi:10.1029/2023MS004080, e2023MS004080 2023MS004080.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S. and Thépaut, J.-N. (2020). The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**(730), 1999–2049, doi:10.1002/qj.3803, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803.

Keisler, R. (2022). Forecasting global weather with graph neural networks. *arXiv preprint arXiv:2202.07575, Feb 15 2022*.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S. and Battaglia, P. (2023). Learning skillful medium-range global weather forecasting. *Science*, **382(6677)**, 1416–1421, doi:10.1126/science.adi2336.

Lang, S., Alexe, M., Chantry, M., Dramsch, J., Pinault, F., Raoult, B., Clare, M. C. A., Lessig, C., Maier-Gerber, M., Magnusson, L., Ben Bouallègue, Z., Prieto Nemesio, A., Dueben, P. D., Brown, A., Pappenberger, F. and Rabier, F. (2024). AIFS — ECMWF's data-driven forecasting system. *arXiv preprint arXiv:2406.01465*, doi:10.48550/arXiv.2406.01465, URL https://arxiv.org/abs/2406.01465.

Lawrence, H., Bormann, N., Sandu, I., Day, J., Farnan, J. and Bauer, P. (2019). Use and impact of Arctic observations in the ECMWF numerical weather prediction system. *Quart. J. Roy. Meteor. Soc.*, **145**(725), 3432–3454, doi:10.1002/qj.3628, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3628.

McNally, A., Lessig, C., Lean, P., Boucher, E., Alexe, M., Pinnington, E., Chantry, M., Lang, S., Burrows, C., Chrust, M., Pinault, F., Villeneuve, E., Bormann, N. and Healy, S. (2024). Data driven weather forecasts trained and initialised directly from observations. *arXiv preprint arXiv:2407.15586*, doi:10.48550/arXiv.2407.15586, URL https://arxiv.org/abs/2407.15586.

Moldovan, G., Pinnington, E., Prieto Nemesio, A., Lang, S., Ben Bouallègue, Z., Dramsch, J., Alexe, M., Santa Cruz, M., Hahner, S., Cook, H., Theissen, H., Clare, M., O'Brien, C., Polster, J., Magnusson, L., Mertes, G., Pinault, F., Raoult, B., de Rosnay, P., Forbes, R. and Chantry, M. (2026). Aifs single 1.1.0: an update to ecmwf's machine-learned weather forecast model aifs. *Geoscientific Model Development*, **19**(10), 4703–4724, doi:10.5194/gmd-19-4703-2026, URL https://gmd.copernicus.org/articles/19/4703/2026/.

Nguyen, T., Shah, R., Bansal, H., Arcomano, T., Madireddy, S., Maulik, R., Kotamarthi, V., Foster, I. and Grover, A. (2023). Scaling transformer neural networks for skillful and reliable medium-range weather forecasting. *arXiv preprint arXiv:2312.03876*.

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z., Azizzadenesheli, K. and Hassanzadeh, P. (2022). Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214, Feb 22 2022*.

Rabier, F., Järvinen, H., Klinker, E., Mahfouf, J.-F. and Simmons, A. (2000). The ECMWF operational implementation of four-dimensional variational assimilation. I: Experimental results with simplified physics. *Quarterly Journal of the Royal Meteorological Society*, **126**(564), 1143–1170, doi:10.1002/qj.49712656415.

WMO (2024). 8th WMO workshop on the impact of various observing systems on numerical weather prediction and earth system prediction (8th impact workshop) - final report. *Technical report*, WMO, URL https://wmoomm.sharepoint.com/:b:/s/wmocpdb/EQKjuBmDcO5GhC0nobUuc8YB6MU7YsrHdtB5JNS0tzMPvg?e=47Xtet.