Exploring Quantitative Observation Impact in Partial and Continuous Cycling Ensemble Kalman Filter Data Assimilation Systems

Gimena Casaretto, [a,b] Craig S. Schwartz, [c] Maria Eugenia Dillon, [a,b] Yanina Garcia Skabar, [a] Juan J. Ruiz, [d,e,f]

[a] *Servicio Meteorologico Nacional, Buenos Aires, Argentina*

[b] *Consejo Nacional de Investigaciones Cientificas y Tenicas (CONICET), Buenos Aires, Argentina*

[c] *National Science Foundation National Center for Atmospheric Research.*

[d] *Centro de Investigaciones del Mar y la Atmósfera (CIMA/CONICET-UBA), Argentina.*

[e] *Instituto Franco-Argentino para el Estudio del Clima y sus Impactos (UMI IFAECI/CNRS-CONICET-UBA), Argentina*

[f] *Departamento de Ciencias de la Atmósfera y los Océanos, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina*

Corresponding author: Gimena Casaretto, gcasaretto@smn.gob.ar

Early Online Release: This preliminary version has been accepted for publication in *Weather and Forecasting*, may be fully cited, and has been assigned DOI 10.1175/WAF-D-24-0127.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

ABSTRACT: This study applies the Ensemble Forecast Sensitivity to Observation Impact (EFSOI) technique to two 80-member ensemble Kalman filter (EnKF) data assimilation (DA) systems over the United States, differing only in cycling strategy: continuous cycling (CC) and partial cycling (PC). EFSOI calculations were performed using 1-hour, 6-hour and 12-hour evaluation forecast times, verified against the Rapid Refresh Model (RAP) analysis. Beneficial impact rates indicated that most observations were beneficial for both DA systems and forecast times, with no significant difference between PC and CC. Differences in cumulative observation impacts were statistically significant only for sources with few observations and small impacts, like mesonet observations. For numerous and impactful observations, such as rawinsondes and aircraft, differences were not statistically significant, suggesting similar use of important observations by PC and CC. PC forecasts were better than CC forecasts, but this improvement is not clearly due to better use of observations. Variable-wise analysis showed similar behavior between PC and CC for impact rates and cumulative impacts of U, V, T, RH, and surface zonal wind. Overall, there was no evidence that either PC or CC systematically used observations better, with mixed results across observation types and sources. Differences between PC and CC were typically small and not statistically significant for the most impactful observations and variables. Fundamental methodological differences between PC and CC did not significantly impact their ability to assimilate observations, the process of ingesting global fields likely responsible for improved PC forecasts relative to CC.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-24-0127.1.

SIGNIFICANCE STATEMENT: This study compares two ensemble Kalman filter (EnKF) data assimilation (DA) systems—continuous cycling (CC) and partial cycling (PC)—using the Ensemble Forecast Sensitivity to Observation Impact (EFSOI) method. Analyzing 1-hour, 6-hour and 12-hour forecasts, most observations were beneficial for both systems, with no significant difference in beneficial impact rates. While PC forecasts were better than CC, this was not clearly due to better observation usage. Significant differences in cumulative observation impacts were only found for less impactful sources. Overall, both systems similarly utilized key observations, and fundamental methodological differences between PC and CC did not significantly affect their observation assimilation capabilities.

1. Introduction

Precise numerical weather prediction (NWP) model forecasts depend on accurate initial conditions provided by data assimilation (DA) systems that combine observations and short-term forecasts. Nowadays, DA systems assimilate a large number of observations from various platforms such as rawinsondes, aircraft, and satellites, which is one of the main reasons forecasts have improved steadily over the last few decades (Bauer et al. 2015). Understanding how observations impact forecasts is critical to improving DA systems and the performance of subsequent forecasts (Sommer and Weissmann 2014; Gustafsson et al. 2018; Hu et al. 2023). Many observation types have a beneficial impact that improve analysis and forecast quality, while others may have detrimental impacts on analysis and forecast quality, pointing to necessary future developments about how these observations are best used (or not used).

There are different methods to calculate the contribution of an individual observation to a DA system, known as observation impact. One method is the ensemble forecast sensitivity to observations impact (EFSOI, Kalnay et al. (2012)). This approach allows quantification of how much each individual observation improved or degraded the forecast, when assimilated with any ensemble Kalman filter (EnKF; Houtekamer and Zhang (2016)), based on the forecast error behavior. The observation impact at the evaluation forecast time is obtained by the difference between the errors in forecasts initialized from two subsequent DA cycles, compared with a verifying truth, which can either be the analysis of the DA system or an independent field (i.e., analysis of another DA system, reanalysis, observations).

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-24-0127.1.

EFSOI has been efficiently used to quantify observation impact in both global DA systems (e.g., Ota et al. (2013); Yamazaki et al. (2021)) and regional DA systems (Kunii et al. 2012; Casaretto et al. 2023). The EFSOI method is also used to monitor the performance of operational DA systems, and recent studies demonstrated that EFSOI information was useful for detecting and rejecting detrimental observations, thereby improving forecasts (Hotta et al. 2017; Chen and Kalnay 2020; Casaretto et al. 2023).

EFSOI can also be applied to investigate how differently constructed EnKFs utilize observations. Specifically, within limited-area DA systems, there are two possible cycling strategies: continuous cycling (CC) and partial cycling (PC). In a CC system, the short-term forecast initialized from the previous analysis serves as the foundation for the current analysis cycle. External models primarily contribute boundary conditions, resulting in a self-contained, limited-area DA system. In contrast, a PC system periodically discards limited-area analysis cycles and replaces them with external analyses generated by a global NWP model. In essence, a PC system periodically incorporates global model information over the whole domain, while a CC system does not. Accordingly, PC and CC DA systems produce different initial conditions and, consequently, diverse forecast outcomes. For instance, Hsiao et al. (2012) demonstrated that compared to CC 3DVAR analyses, PC 3DVAR analyses had substantially smaller biases and initialized notably improved forecasts over Taiwan and its surrounding areas. Furthermore, numerous studies employing CC DA over the conterminous United States (CONUS) and adjacent regions have documented bias accumulations (e.g., Torn and Davis (2012); Romine et al. (2013); Cavallo et al. (2016); Poterjoy et al. (2021)).

Given their different cycling methodologies, the impacts of specific observation types may vary considerably between PC and CC systems. Understanding how PC and CC DA systems utilize observations is crucial, considering that both are used operationally. For example, the KENDA DA system (Schraff et al. 2016) employs continuous cycling, whereas other systems like the operational RAP (Benjamin et al. 2016) and HRRR (Dowell et al. 2022) over the CONUS use partial cycling. Thus, to better understand how PC and CC DA systems use observations, this study employs EFSOI to explore observation impact in both CC and PC 80-member, 15-km EnKFs over the CONUS that were identical except for cycling strategy. Using EFSOI to compare both CC and PC DA systems can provide unique insights as to how these different types of DA systems use observations. Furthermore, providing objective guidance on the impact of specific observation types within

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-24-0127.1.

limited-area DA systems will help set expectations for future field campaigns. Moreover, this research aims to enhance the use of observations provided by various observing systems and inform the configurations of future DA systems.

This article is structured as follows: section 2 describes the DA systems, model configurations, and the EFSOI algorithm. In section 3, we analyze both the forecasts' behavior relative to the verifying truth used in the EFSOI algorithm and the EFSOI results. Finally, section 4 provides a summary and conclusion of this study.

2. Methodology

a. DA systems

EFSOI was applied to the CC and PC EnKFs described in Schwartz et al. (2022); hereafter S22. Specifically, the PC and CC EnKFs both had 80 ensemble members with 15-km horizontal grid spacing and were executed over the CONUS and adjacent areas (Fig. 1). Both EnKFs performed hourly DA cycles over a period spanning 23 April to 20 May 2017 and used version 3.9.1.1 of the Advanced Research Weather Research and Forecasting (WRF) model (Skamarock et al. 2008; Powers 2017). A serial EnKF from the Data Assimilation Research Testbed (Anderson et al. 2009) was used in both the PC and CC EnKFs.

In the CC EnKF, after initializing the EnKF from a global model at 0000 UTC 23 April 2017, the background for DA was always the 1-hour ensemble forecast initialized from the previous analysis. Thus, after 2 days of continuous DA cycles, the CC EnKF lost "memory" of its initialization from the global model (S22), and other than lateral boundary conditions, the CC EnKF was independent of global fields. Conversely, in the PC EnKF, daily at 1200 UTC, the EnKF was initialized from a global ensemble (Fig. 2). Then, 12 hours of hourly forecast-analysis cycles were produced until 0000 UTC, after which, DA cycles ceased (i.e., the PC EnKF did not perform analyses between 0100 UTC – 1100 UTC). The next day, the PC EnKF was again initialized from a global ensemble (S22). Aside from this difference in cycling methodology, both the CC and PC EnKFs had identical configurations. Because 12 hours of cycling was performed in the PC EnKF, by 0000 UTC, the PC EnKF had spun-up somewhat, but S22 showed how large scales in PC EnKF analyses still had "memory" of the global ensembles that were used daily to initialize the PC EnKF. More details are found in S22.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-24-0127.1.

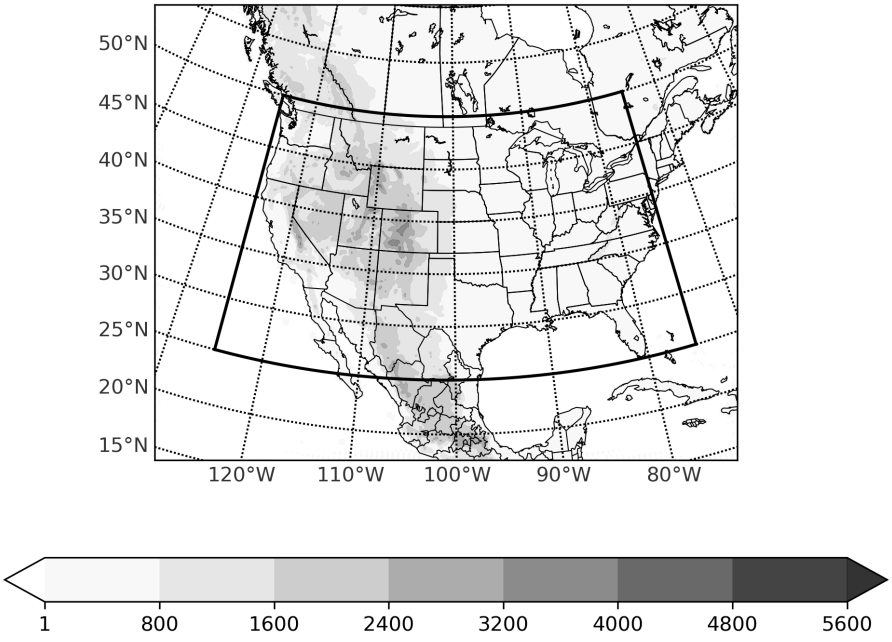

FIG. 1: Computational domain with 15-km horizontal grid spacing, the corresponding model topography in meters (shaded), and EFSOI target area (solid line).

As EFSOI requires forecasts initialized from adjacent analyses, the analyses from the CC and PC EnKFs at both 2300 UTC and 0000 UTC were used to initialize 15-hour 80-member ensemble forecasts. The output of these 15-hour forecasts was then used in EFSOI calculations (a total of 23 forecasts). Although this work used the same DA systems as S22, the forecasts produced and analyzed here fundamentally differed from S22. Specifically, S22 focused on 10-member ensemble forecasts with 3-km horizontal grid spacing over the CONUS (where the 3-km forecasts were initialized by downscaling 15-km EnKF analyses) whereas here we focus on 80-member, 15-km ensemble forecasts as a first approach to quantitatively investigate the observation impact in PC and CC DA systems.

The observations assimilated in both EnKFs included data from rawinsondes, aircraft, wind profilers, global positioning system radio occultation (GPSRO), infrared and water vapor channel satellite-tracked wind, marine (including ship and buoy), land surface (including METAR and SYNOP), and the Oklahoma and West Texas mesonet observing platforms (Table 2 in S22). Identical observations were introduced into both the CC and PC EnKFs for assimilation. However, an "outlier check" was applied as part of quality control, where observations were rejected if the ensemble mean innovation was too large. Thus, in reality, the PC and CC EnKFs assimilated

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-24-0127.1.

slightly different observations: CC assimilated 2261277 observations at the 0000 UTC cycle over the entire experimental period while PC assimilated 2263233 observations; i.e. PC assimilated 0.08% more observations than CC.

b. EFSOI method

EFSOI calculations were performed within the EFSOI target area (Fig. 1) for both the CC and PC EnKFs, following Kalnay et al. (2012). Instead of using their own analyses for EFSOI calculations, an independent verifying truth was used, analyses from the ~ 13-km Rapid Refresh Model (RAP; Benjamin et al. (2016)), which were interpolated to the 15-km model grid using the WRF preprocessing system to compute the calculations. RAP analyses were chosen as the verifying truth to ensure that both DA systems' errors were measured against the same reference. Moreover, as Kotsuki et al. (2019) suggested, the use of a generic independent analysis can avoid EFSOI overestimations of the beneficial impact rate of observations.

EFSOI was computed for 1-hour, 6-hour and 12-hour evaluation forecast times. It is worth considering that a longer evaluation forecast time likely makes the EFSOI computation less accurate due to linearization and localization advection errors. The schematic formulation of the EFSOI applied to both DA systems is represented in Figure 2.

The EFSOI methodology employed is the same as in Casaretto et al. (2023), and the following text is derived from there with minor modifications. The change in the forecast error ($e_{t|23} = X_{t|23} - X_t^v$ and $e_{t|00} = X_{t|00} - X_t^v$, Fig. 2) produced by the assimilation of observations at 0000 UTC is approximately given by

$$\triangle e^2 \cong \frac{1}{K-1} \delta y_{00}^T [\rho \circ R^{-1} Y^a X_{t|00}^{fT}] C (e_{t|00} + e_{t|23}) \tag{1}$$

where K is the ensemble size, R is the observation error covariance matrix, $Y^a = HX^a$ is the matrix of analysis ensemble perturbations in observation space valid at time 00 UTC, $X_{t|00}^{fT}$ is the ensemble forecast perturbation valid at time T=t initialized at time 00 UTC, and δy_{00} is the observation minus background (O-B) innovation vector at time 00 UTC. To reduce the impact of sampling noise in covariance estimates due to the use of a limited size ensemble, the EnKFs applied localization (horizontal localization full width of 1280 km and vertical localization full width of 1.0 scale height; S22). Localization consists of the element wise multiplication of the

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-24-0127.1.

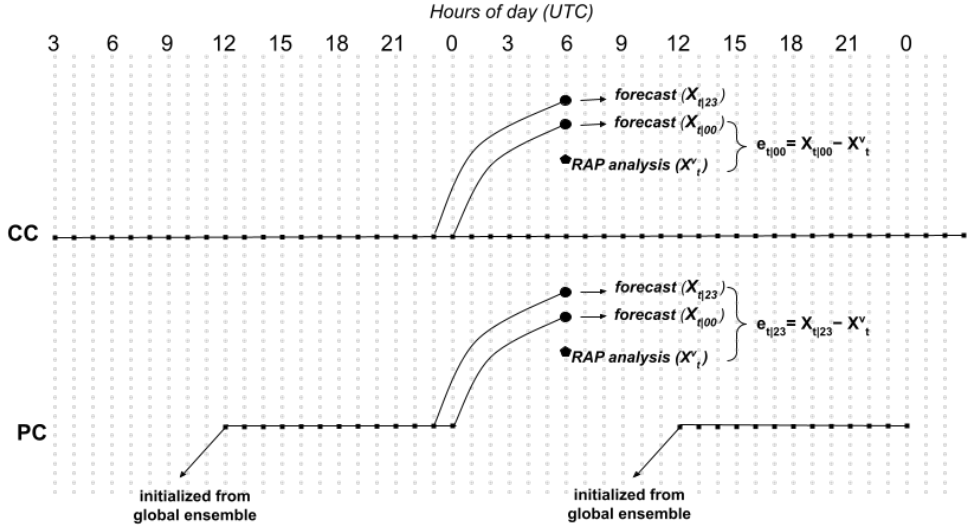

FIG. 2: Schematic diagram of the implementation of EFSOI in CC and PC DA systems. The background of CC EnKF is always the 1-hour ensemble forecast initialized from the previous cycle's analysis, while the PC EnKF background initializes from a global ensemble daily at 1200 UTC. Black squares represent EnKF analyses. The forecast error of the forecast initialized at 2300 UTC ($t|23$) and 0000 UTC ($t|00$) are calculated as the difference between these forecasts and the RAP analysis (X_t^v ; verifying truth) at the evaluation forecast time t ($e_{t|23}$ and $e_{t|00}$) for both DA systems. The difference between ($t|23$) and ($t|00$) is due to the assimilation of observations at the target analysis time (which is 0000 UTC in this case).

matrix $R^{-1}Y_a X_{t|00}^{fT}$ by the localization matrix . It is assumed that the evaluation forecast time is short enough that the impact of advection upon localization can be neglected (Lien et al. 2018).

In [Eq. 1], C is a positive definite matrix to measure the error of the forecast, which is necessary for the EFSOI formulation but is not calculated explicitly. In this study we used the moist total energy norm (MTE, Ehrendorfer et al. (1999)):

$$e^T C e = MTE = \frac{1}{2}\frac{1}{S}\int_S \left[\int_0^1 (u'^2 + v'^2 + \frac{C_p}{T_r} T'^2 + \frac{L^2}{C_p T_r} q'^2)\, d\sigma + \frac{R_d T_r}{P_r^2} P_s'^2 \right] dS \qquad (2)$$

The difference between the state of the model and the reference (i.e., verifying truth) is expressed as primed variables (u',v',T',q',P') of the equation. The following constants were considered: T_r = 280 K (reference temperature), c_p = 1,012 J $g^{-1}K^{-1}$ (specific heat of the air at constant pressure), L = 2510400 J Kg^{-1} (latent heat of condensation per unit mass), R_d = 287 J Kg^{-1} K^{-1} (dry air constant), and P_r = 10^5 Pa (reference pressure). The integration extends over the full EFSOI horizontal domain (S, Fig. 1) and vertical domain (levels, 1000hPa to 100 hPa).

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-24-0127.1.

If the observation impact [Eq. 1] from assimilating a given observation is negative, that indicates a reduction in forecast error, whereas a positive observation impact value indicates an increase in forecast error. Each observation is assigned an impact value for each DA cycle within the EFSOI target area (Fig. 1). Based on this information, different statistics can be computed over the entire observation dataset or over different subsets, such as a particular observation source (e.g., rawinsondes) or a particular observed quantity (e.g., pressure, temperature).

3. Results

a. Forecast comparison to RAP analyses

The forecasts corresponding to evaluation forecast times were compared to RAP analyses, which also served as the verifying truth for the EFSOI in this study. This analysis was done to examine how apart the two DA systems were and to correlate these results with EFSOI calculations. RAP analyses were interpolated to the 15-km model grid using the WRF preprocessing system to compute various statistics, and the comparison was conducted in the EFSOI target area (Fig. 1).

The time- and horizontally-averaged root-mean-square error (RMSE) and bias (forecast minus analysis) of the ensemble mean forecast with respect to RAP analyses for zonal (U) and meridional (V) wind components, temperature (T), and specific humidity (Q) are shown in Figure 3. Generally, RMSE differences among the DA systems are larger for U and V compared to the thermodynamic variables (T and Q), with PC having smaller RMSEs than CC for almost all pressure levels and both forecast times (Figs. 3 c,d,g,h). For T and Q, the largest RMSE differences are in the lower levels and PC has the smallest RMSEs, while above 500 hPa both DA systems have similar RMSEs (Figs. 3 a,b,e,f). For U, V, and T, RMSE differences between CC and PC are larger at the 1-hour lead time than at the 6-hour lead time, while for Q, differences between CC and PC are similar at both forecast lead times. Differences between PC and CC RMSEs may be smaller at 6-hour than at 1-hour because the impact of initial conditions decreases as the forecast lead time increases.

Biases of T, U, and V differ for each level and forecast lead time, and there is no clear signal of which DA system performs best. The bias of Q is smaller for PC than for CC close to the surface, 1000 hPa up to 500 hPa (Figs. 3b,f). RMSEs and biases for the 12-hour evaluation forecast time yield similar results as those for the 6-hour lead time, although the PC and CC values are even closer than at 6-hour (not shown).

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-24-0127.1.

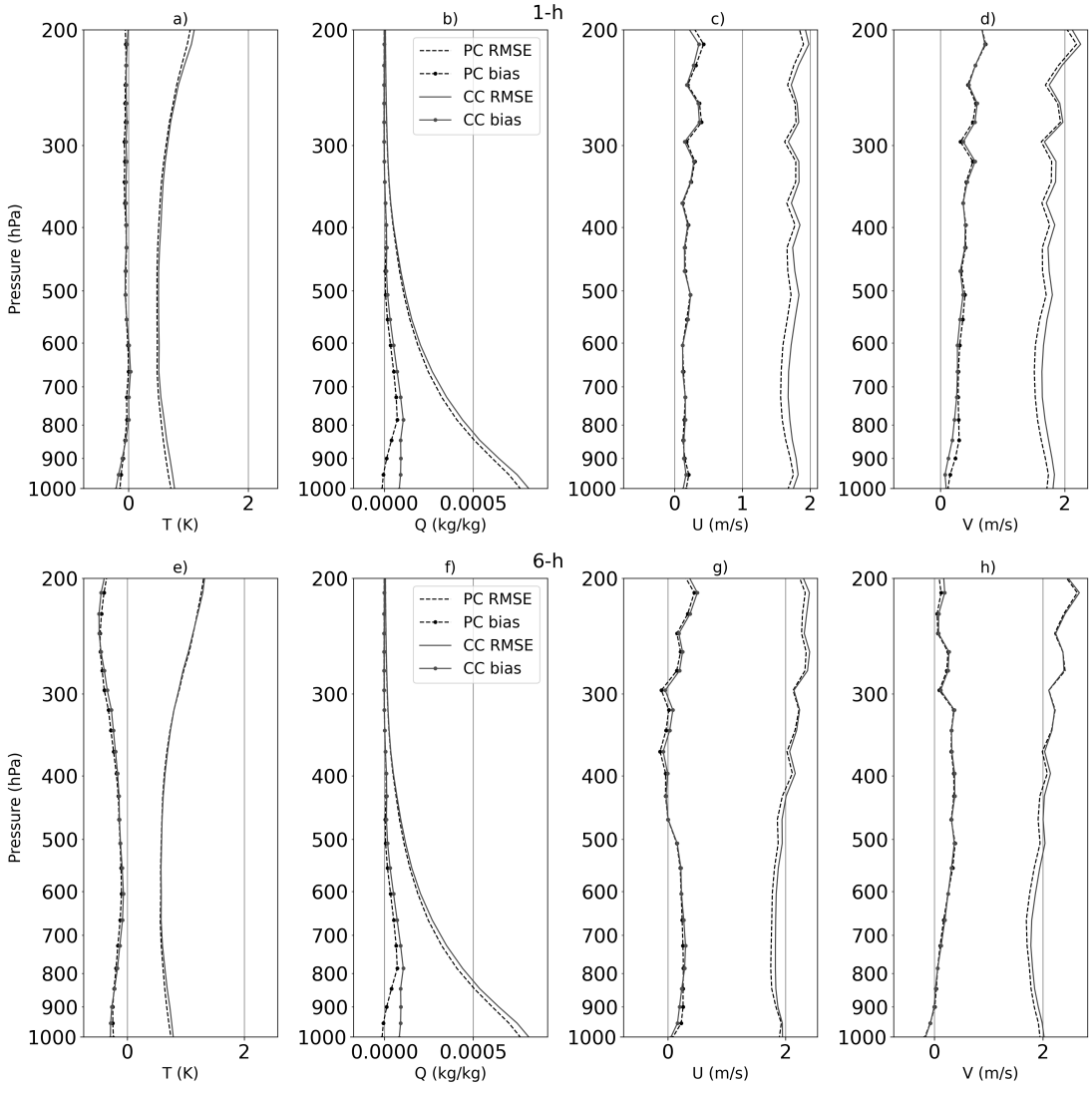

FIG. 3: Vertical profiles of the temporally and spatially averaged bias (forecast minus analysis; lines with circle markers) and RMSE (lines without markers) of PC (dashed lines) and CC (solid lines) ensemble mean forecasts with respect to RAP analyses. a), e) Temperature [K]; b),f) Specific humidity [$KgKg^{-1}$]; c),g) zonal wind component [ms^{-1}] and d),h) meridional wind component [ms^{-1}]. a)-d) is for the 1-hour forecast lead time and e)- h) is for the 6-hour forecast lead time for forecasts initialized at 0000 UTC.

To ensure that these results are not biased because the RAP uses a partial cycling DA system (Benjamin et al. 2016), the same analysis was carried out by comparing the PC and CC forecasts with the European Centre for Medium-Range Weather Forecasts fifth-generation atmospheric Reanalysis (ERA5, Hersbach et al. (2020)). The same behavior comparing the DA systems was noted: PC has smaller RMSEs than CC (not shown).

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-24-0127.1.

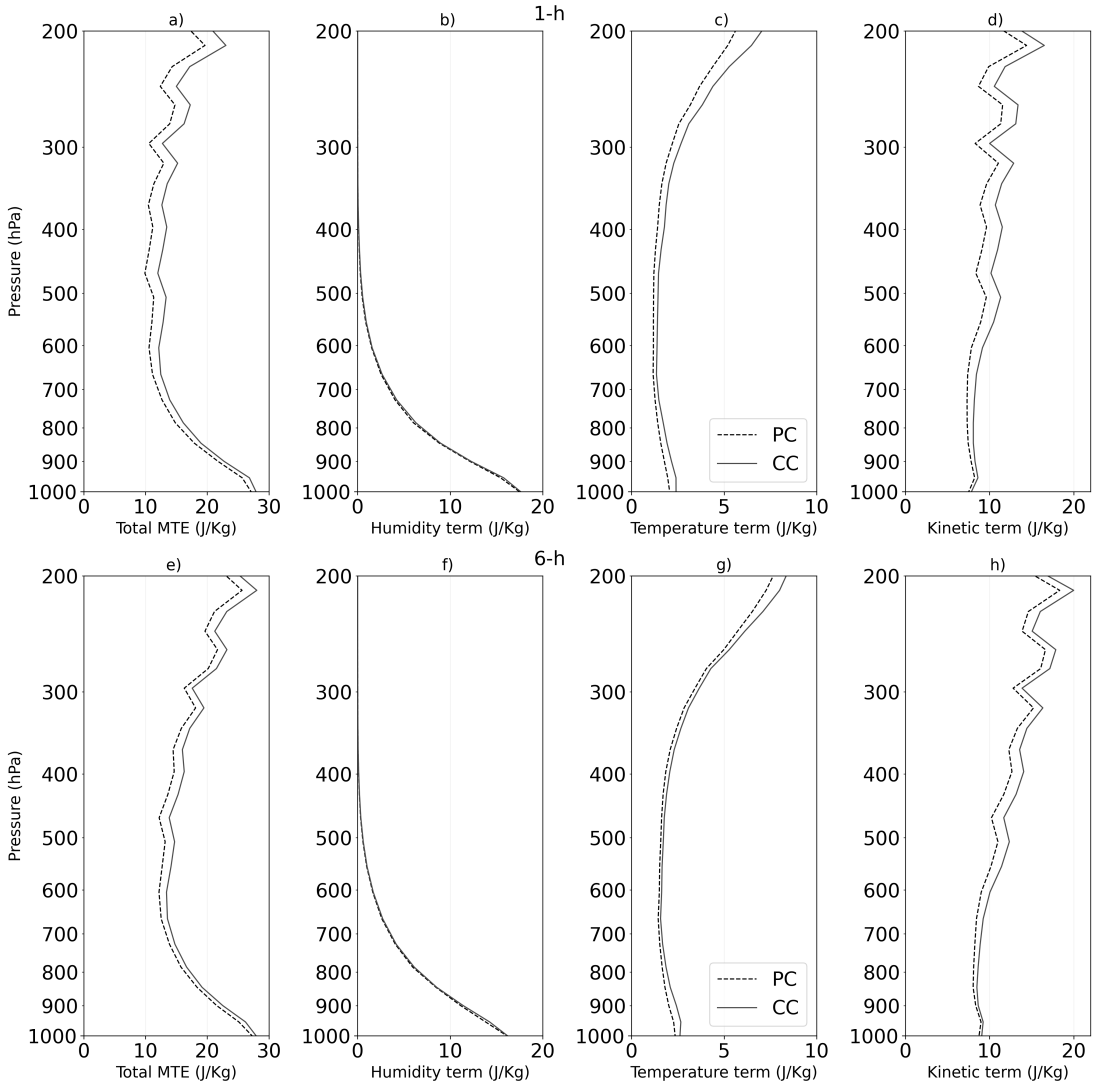

FIG. 4: Vertical profiles of the temporally and spatially averaged MTE and the terms that contribute to the MTE [JKg^{-1}], calculated for PC (dashed line) and CC (solid line) ensemble mean forecasts with respect to RAP analyses. a),e) moist total energy; b),f) humidity term; c),g) temperature term; and d),h) kinetic term. a)-d) 1-hour forecast lead time and e)-h) 6-hour forecast lead time for forecasts initialized at 0000 UTC.

Figure 4 displays the vertical profiles of MTE and each term related to kinetic energy, temperature, and humidity [e.g., Eq. 2] for the ensemble mean forecasts. The RAP analysis is used as a reference to ensure comparable results to the terms of the EFSOI equation. The term related to surface pressure is not shown, as its contribution is between one and two orders of magnitude smaller.

For total MTE and the kinetic term, the errors at 1-hour forecast lead time (Figs. 4a, d) show larger differences between CC and PC than the 6-hour forecast lead time (Figs. 4e, h). The kinetic

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-24-0127.1.

term reveals that PC has lower error values compared to CC, indicating better performance when using RAP as a reference, for both 1-hour and 6-hour forecast lead times (Figs. 4e, h). These overall results align with the RMSEs in Fig. 3. The humidity term of both DA systems have similar errors at most levels (Figs. 4b,f), except for the lower levels (1000-900 hPa) where PC outperforms CC. Compared to the humidity term, the temperature term differences are more substantial and PC outperforms CC at all levels and forecast times (Figs. 4c,g). Lastly, the total MTE for both forecast times is smaller for PC than for CC through the entire profile, indicating that it is dominated by the kinetic term in mid and upper levels and by the temperature term between 1000 and 800 hPa (Figs. 4a,e). In addition, the humidity term seems to be the term that most contributes to the maximum of MTE near the surface for both DA systems. MTE evaluated at the 12-hour lead time again provides similar results as at the 6-hour lead time, although differences between PC and CC are smaller (not shown).

Differences in forecast performance between PC and CC could be captured by examining the EFSOI calculations, which will be presented in section 3.b-d. Given that PC generally has a more accurate background state than CC (e.g., Figs. 3, 4), observations in CC could contribute a relatively higher impact, highlighting areas where forecasts benefit more substantially from data assimilation. These variations in observation impact are likely influenced by the distinct forecast error dynamics in PC and CC, where observations help to adjust and improve the less accurate CC background state. Thus, subtle but meaningful differences in observation impact between PC and CC may be revealed through EFSOI. Understanding how observation impacts vary between PC and CC systems is essential for optimizing data assimilation strategies. By analyzing EFSOI with a focus on these forecast performance differences, we can improve our approach to integrating observational data, ultimately enhancing forecast accuracy in models with varying background states.

The EFSOI definition in [Eq. 1] states that the total impact of all assimilated observations is represented by the difference in forecast error between two consecutive forecasts. The difference between the MTE error of forecasts initialized at 0000 UTC and the MTE error of forecasts initialized at 2300 UTC, both valid at the same time and using the RAP analysis as a reference, can be interpreted as the total impact of the assimilated observations. Therefore, the MTE error was calculated for each grid point and then was averaged to get a single value representative for

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-24-0127.1.

the EFSOI target area for each DA system, both for the forecasts initialized at 00 UTC (MTE_00) and for the forecasts initialized at 23 UTC (MTE_23), for the valid times 01, 06 and 12 UTC. Afterwards, these single values were subtracted to get the difference MTE_diff defined as MTE_00 minus MTE_23.

	Valid forecast time		
	01 UTC	*06 UTC*	*12 UTC*
MTE_diff PC $[JKg^{-1}]$	-0.4144	-0.7427	-0.9614
MTE_diff CC $[JKg^{-1}]$	-0.4022	-0.7589	-0.9533
$\lvert$*[MTE_diff PC] - [MTE_diff CC]*$\rvert$ $[JKg^{-1}]$	0.0115	0.0162	0.0029

TABLE 1: Difference between the MTE error of forecasts initialized at 0000 UTC and the MTE error of forecasts initialized at 2300 UTC, both valid at the same lead time and using the RAP analysis as a reference, for each DA system (MTE_diff PC, MTE_diff CC), for three valid times. The difference between the DA systems is also shown. See the text for more details.

Table 1 shows the results of MTE_diff for the three valid evaluation times, which are the same evaluation forecast times used for EFSOI. The negative values of MTE_diff for both DA systems and across all the valid times indicate that the assimilation of observations at 0000 UTC reduces the forecast errors. As lead time increases, the differences between PC and CC diminish (i.e. [MTE_diff PC] - [MTE_diff CC]), while the overall error reduction grows. The observation impact differences between PC and CC may be more pronounced at shorter forecast times, as the forecasts may contain more information from the global ensemble in the background. Additionally, as the lead time increases, forecast errors such as MTE, bias, and RMSE become more similar between PC and CC, as shown previously in Figures 3 and 4. Depending on the lead time, observations may reduce forecast errors to a greater extent in either PC or CC (Table 1).

b. EFSOI results per observation sources

Based on the EFSOI computation described in section 2.b., we calculated both the cumulative observation impact and the beneficial impact rate for each observation source to identify variations in the influence of observations within the PC and CC DA systems (Fig. 5). The cumulative observation impact represents the sum of all individual observation impacts for a group of observations (in this case, categorized by observation source).

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-24-0127.1.

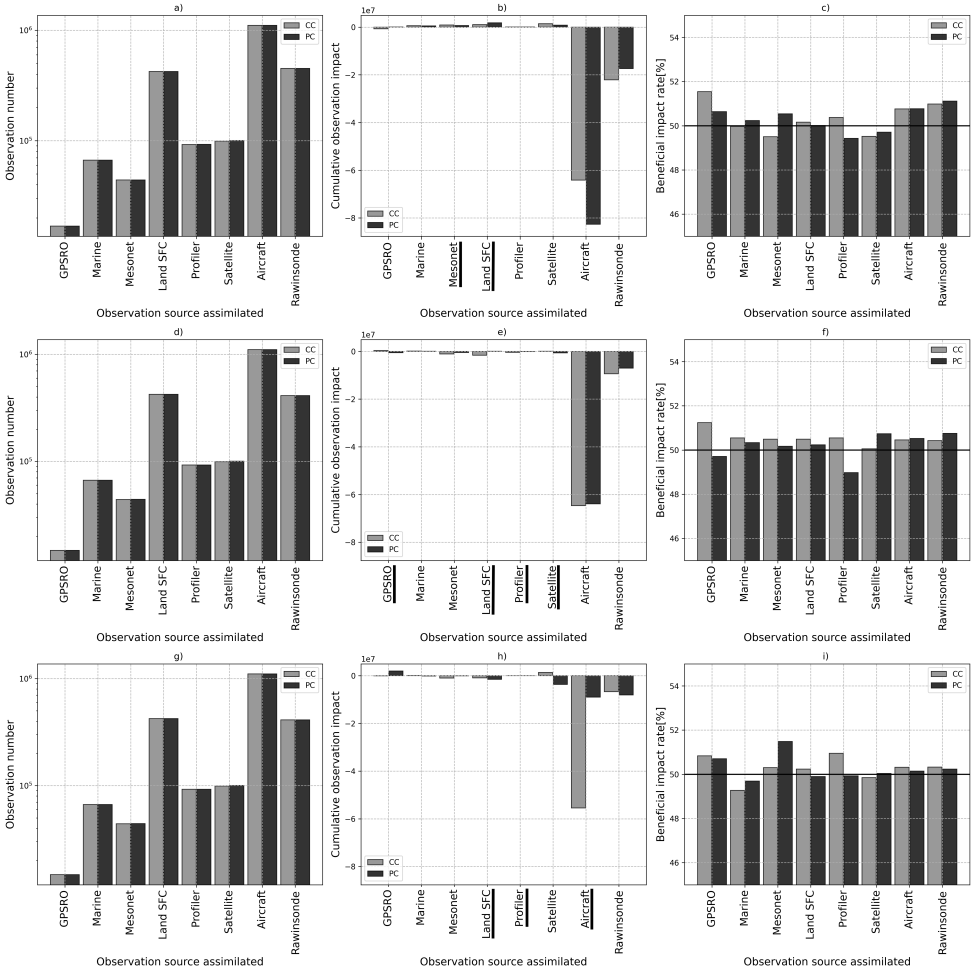

FIG. 5: a),d), g) Observation number (logarithmic scale); b),e), h) cumulative observation impact [JKg^{-1}] and c),f),i) beneficial impact rate [%] per observation source assimilated, for CC and PC over all 23 cases. a)-c) 1-hour , d)-f) 6-hour g)-i) 12-hour forecast lead time for forecasts initialized at 0000 UTC. Underlined observation sources have statistically significant differences with 95% confidence based on the Mann-Whitney-Wilcoxon rank-sum test comparing CC and PC trimmed distributions between the 1st and 99th percentiles of observation impact [JKg^{-1}].

First, as the evaluation forecast time increases, the cumulative beneficial impact and beneficial impact rate both decrease (Fig. 5b,e,h and c,f,i). This pattern aligns with findings from previous studies where EFSOI overestimates, especially for shorter lead times such as 6-hour (Kotsuki et al. 2019). The overall impact of the observation for both DA systems and evaluation forecast times was beneficial and aligns with general findings from past studies (Lorenc and Marriott 2014; Sommer and Weissmann 2014; Cardinali 2018), which reported beneficial impact rates between 50% and 55%. Both CC and PC have comparable proportions of beneficial observations for all the evaluation forecast times.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-24-0127.1.

Despite differences in forecast performance (section 3.a.), these discrepancies do not appear to be explained by variations in observation impact. When comparing the DA systems (CC and PC) with EFSOI, similar results are obtained across some subsets. For instance, aircraft and rawinsondes—two observation sources with the highest numbers of assimilated observations (Figs. 5a, d, g)— exhibit both substantial cumulative observation impact and beneficial impact rates above 50%. This results in a consistently beneficial impact across the three forecast times and DA systems (Figs. 5b, c, e, f, h, i). Moreover, the cumulative impact values stand out significantly compared to other subsets. These findings align with James and Benjamin (2017), who also highlighted aircraft observations as a significant contributor to the overall impact of observations.

To understand observation impact, we also examined distributions of the observation impact results of the subsets of the observations assimilated. They revealed many outliers, so the data was trimmed between the 1st and 99th percentiles and the Mann–Whitney–Wilcoxon rank-sum test (MWW, Wilks (2011)) was used to assess whether differences between CC and PC trimmed distributions were statistically significant. The MWW test is a non-parametric statistical test used to determine whether there is a significant difference between the distributions of two independent samples. The test works by ranking all the values from both groups together and then comparing the sums of these ranks. The null hypothesis is that the distributions of the two groups are identical while the alternative states that the distributions of the two groups are different.

According to the MWW rank-sum test, rawinsonde observations impacts in PC and CC were not statistically significantly different, indicating that both DA systems used these observations similarly (Figs. 5b,e,h). Aircraft observations were only statistically different for 12-hour forecast lead time, where the cumulative observation impact between CC and PC is broader (Fig. 5h). Additionally, the estimated impacts decrease as lead time increases, likely due to the limitations of the localization used in EFSOI, which struggles to account for time evolution, causing observation impacts to diminish over time. This limitation is more pronounced in the upper troposphere, where faster wind flows advect the impacts beyond the localization radii. As a result, the impacts from aircraft observations likely decay more quickly than those from surface observations. Hotta et al. (2017) encountered similar results regarding observation impact and the evaluation forecast lead time.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-24-0127.1.

Results for Land SFC observations are statistically significantly different for the forecast lead times and is the only subset with this behavior. The cumulative observation impact varies across the forecast lead times; for the shorter lead times the observations are detrimental (positive value) while they are beneficial for the longer times (negative value). Regarding the beneficial impact rate, the values remain around 50% for all the forecast lead times.

The marine source exhibits a detrimental cumulative impact for PC and CC at 1-hour and 6-hour forecast lead times (Figs. 5b,e), but beneficial impact rates have values $\geq 50\%$ (Figs. 5c,f). This finding means that while some marine observations may have a large detrimental impact, when considering the beneficial impact rate—a measure of the relative relationship between detrimental and beneficial observations—more observations are beneficial than detrimental. For the 12-hour forecast lead time, the cumulative observation impacts have opposite signs when comparing PC and CC (Fig. 5h), but for both DA systems, the beneficial impact rate is $< 50\%$ (Fig. 5i), which means that more than half of the observations result in detrimental impacts.

Profiler and mesonet observations have a cumulative detrimental impact for both PC and CC at 1-hour (Fig. 5b), while the opposite occurs at 6-hour and 12-hour (Fig. 5e, h), with mixed results in terms of beneficial impact rates (Figs. 5 c,f,i). For 6-hour forecasts, assimilating satellite-tracked winds clearly improves PC, with a negative cumulative impact in Figure 5e,f. The MWW rank-sum test revealed that these observation impacts were statistically significantly different between CC and PC.

For GPSRO observations, while CC shows a consistent beneficial impact in 1-hour forecasts (negative cumulative impact and beneficial impact rate $\geq 50\%$), the behavior of PC is not consistent (positive cumulative impact and beneficial impact rate $\geq 50\%$) (Figs. 5b,c). However, according to the MWW rank-sum test, PC and CC have no significant differences in the impact of GPSRO observations, and therefore, extreme values of PC might be affecting its positive cumulative impact (Fig. 5b).

These results emphasize the importance of carefully tuning assimilation settings for different observation types, which is essential for optimizing DA systems. Depending on the specific purpose of a DA system—whether solely to produce analyses or also to initialize forecasts from those analyses—adjustments to assimilation techniques for certain observation sources may be warranted. For instance, sensitivity experiments could help determine the effectiveness of including

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-24-0127.1.

or excluding a particular observation source, adjusting localization settings, or applying techniques such as thinning or superobbing to enhance data exploitation. The mixed EFSOI results suggest that factors beyond observational data—possibly related to the DA system configuration itself—could be contributing to the performance differences observed between CC and PC.

Moreover, differences between PC and CC observation impacts were only statistically significant for observation sources providing relatively few observations and small impacts (e.g., mesonet observations). For more numerous and impactful observations, like rawinsondes and aircraft, differences between PC and CC observation impacts were not statistically significant. These findings suggest that PC and CC used these most important observations similarly, and there were mixed results regarding which DA system used the less impactful observations most beneficially. Thus, overall, while PC forecasts were better than CC forecasts (Figs. 3,4), this result does not obviously appear to be attributable to PC making better use of observations than CC. Although the PC EnKF periodically ingests global data (while the CC EnKF does not), by 0000 UTC, the primary impact of the ingested global fields in the PC EnKF is on the large scales (S22). However, observations sample many different spatial scales, meaning that large scales alone are not reflected in EFSOI statistics. Accordingly, different large-scale representations between PC and CC appear to have been insufficient to meaningfully alter observation impacts. Yet, the large-scale state-space differences between the two EnKFs may have had a major role in engendering forecast differences between PC and CC.

c. EFSOI results per observed variables

Continuing our analysis of the diverse impacts of observations across the different DA systems, Figure 6 presents the EFSOI observation impact and beneficial impact rate for the assimilated observed variables, analyzed separately. Consistent with the patterns observed in Figure 5, both the cumulative beneficial impact and the beneficial impact rate decrease as the evaluation forecast time increases (Figs. 6b, e, h, and Figs. 6c, f, i).

The 3D variables—zonal wind (U), meridional wind (V), temperature (T), and relative humidity (RH)—along with the surface zonal wind (US), consistently demonstrate forecast improvements in terms of both cumulative observation impact and beneficial impact rates across the forecast times and DA systems (Figs. 6b, c, e, f, h, i). For specific variables, such as T at 1-hour and 6-hour

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-24-0127.1.

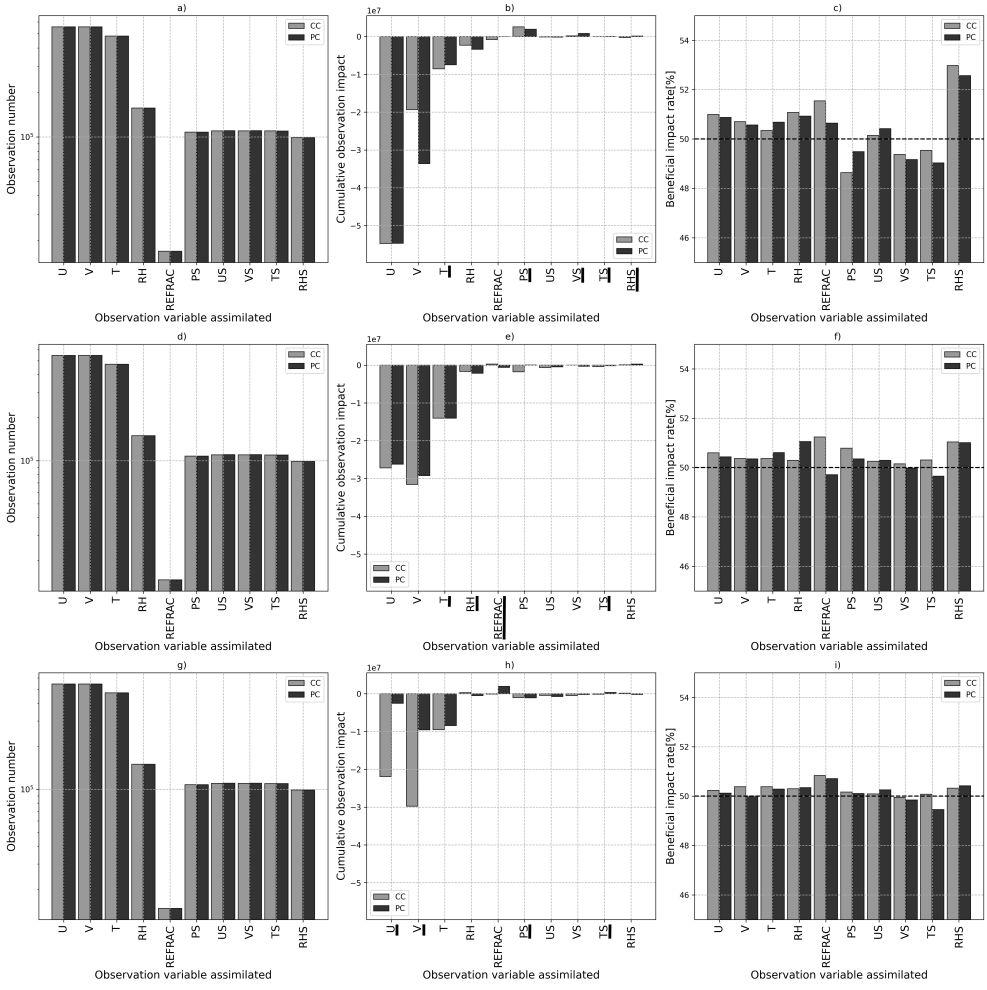

FIG. 6: a),d),g) Observation number (logarithmic scale), b),e),h) cumulative observation impact [JKg^{-1}] and c),f),i) beneficial impact rate [%] per observation variable assimilated, for CC and PC over all 23 cases. a)-c) 1-hour , d)-f) 6-hour g)-i) 12-hour forecast lead time for forecasts initialized at 0000 UTC. Statistics are shown for zonal (U) and meridional (V) wind components; temperature (T); relative humidity (RH), refractivity (REFRAC); surface pressure (PS); surface zonal (US) and meridional (VS) wind components; surface temperature (TS); and surface relative humidity (RHS). Underlined observation variable has statistically significant differences with 95% confidence based on the Mann-Whitney-Wilcoxon rank-sum test comparing CC and PC trimmed distributions between the 1st and 99th percentiles of observation impact [JKg^{-1}].

forecasts, RH at 6-hour forecasts, and the wind components (U and V) at 12-hour forecasts, the differences in cumulative observation impacts between the CC and PC systems are statistically significant, as determined by the MWW rank-sum test. These statistically significant differences may explain the larger variations in beneficial impact rates for these variables compared to others.

Furthermore, surface pressure (PS), meridional wind (VS), and temperature (TS) lead to more positive impacts at 6-hour and 12-hour compared to 1-hour in terms of beneficial impact rate in both

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-24-0127.1.

DA systems (Figs. 6c,f,i). These surface variables exhibit detrimental impacts on the background (Figs. 6b,c), and also larger innovations than other variables when observations are sourced from mesonet, land SFC, and marine platforms (not shown). In contrast, surface relative humidity (RHS) and US have beneficial impact rates $\geq 50\%$ at the three forecast times (Figs. 6c,f,i). Therefore, to improve the background of both DA systems, the assimilation of PS, VS, and TS observations from mesonet, land SFC, and marine data sources should be revised, because more than half of these observations result in a detrimental impact.

Notably, among the variables, U, V, and T have a higher amount of observations assimilated, and they have larger cumulative observation impacts than the remaining variables, indicating a broad impact in both DA systems. Differences between CC and PC are generally small for the wind components for 1-hour and 6-hour forecast lead times, and there is no statistically significant difference among their observation impacts according to the MWW rank-sum test (Figs. 6b,e). However, for 12-hour forecast lead time the differences are bigger and statistically significant for U and V (Fig. 6h), in consonance with the results encountered for aircraft (Fig. 5h). The reasons for these results are unclear. CC has higher beneficial impact rates than PC for U and V for all the forecast lead times (Figs. 6c,f,i). Observation impacts for T statistically significantly differ between PC and CC at 1-hour and 6-hour forecast lead times (Figs. 6b,e), and PC has higher beneficial impact rates for T than CC (Figs. 6c,f). As noted in Section 3.b, these variables also illustrate that the estimated impacts diminish as lead time increases.

There is an overall trend that surface observations are more likely to have detrimental impacts, and upper-air observations are more consistently beneficial. Surface observations are affected by a multitude of local factors, such as terrain, land-use variability, and human activities, which can introduce noise and non-representative data. These local factors make it challenging for DA systems to extract useful information that benefits the entire atmospheric model. Errors or biases in surface observations may propagate into the model, potentially destabilizing it or leading to inaccuracies, hence the detrimental effect. Also, surface observations may not align well with the model's vertical grid, resulting in representativeness errors where the data mismatch the modeled values. In addition, the model itself has difficulties to properly simulate the planetary boundary layer, worsening with this mismatch.

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-24-0127.1.

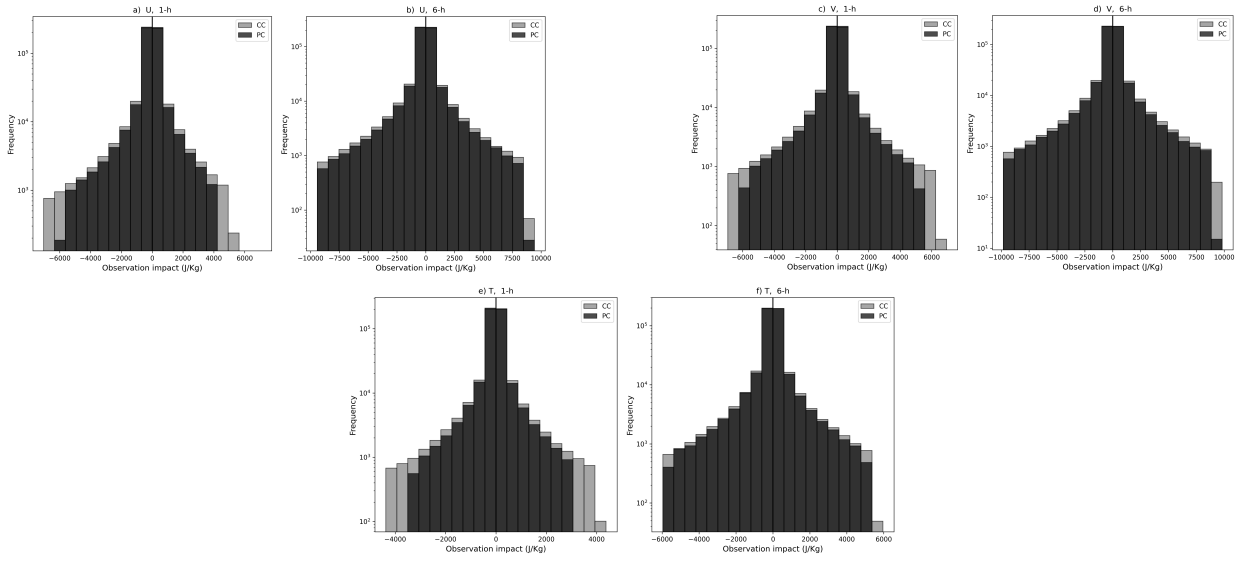

FIG. 7: Trimmed distributions between the 1st and 99th percentiles of observation impact (JKg^{-1}) for both DA systems (CC grey, PC black) over all 23 cases for (a),(b) zonal wind component (U), (c),(d) meridional wind component (V), and (e),(f) temperature (T) for a),c),e) 1-hour and b),d),f) 6-hour evaluation forecast times.

Holistically, these results suggest that for the most numerous and impactful variables, which are the upper air observations (U, V, and T), CC and PC performed similarly. Results for other less numerous and impactful variables were generally mixed, sometimes favoring CC and other times PC. Accordingly, these findings are similar to those provided by Fig. 5 and suggest that poorer CC forecasts relative to PC forecasts (Figs. 3,4) are not clearly due to inferior use of observations in CC relative to PC.

d. Further analysis of EFSOI statistics

The cumulative observation impacts (Figs. 5,6) only provide a general perspective about how the PC and CC EnKFs used observations. To gain a better understanding of observation impact, we will delve deeper into analyzing the distributions of EFSOI statistics. This approach is novel, as previous EFSOI studies have only focused on the cumulative or mean statistics, and examining distributions allows for a deeper assessment of observation impact. As mentioned above, distributions of observation impact revealed many outliers, so the data were trimmed between the 1st and 99th percentiles for mean, median, and standard deviation calculations. In addition, significance tests were performed to determine if differences between the means, medians, and standard deviations

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-24-0127.1.

	1-h			6-h		
	Mean	Median	STD	Mean	Median	STD
U	CC <PC		CC >PC			CC >PC
V			CC >PC			CC >PC
T		CC >PC	CC >PC		CC >PC	CC >PC
U-aircraft	CC <PC		CC >PC			CC >PC
U-rawinsonde	CC <PC		CC >PC			CC >PC
U-profiler	CC >PC		CC >PC	CC >PC	CC <PC	CC >PC
U-satellite			CC >PC			
V-aircraft			CC >PC			CC >PC
V-rawinsonde			CC >PC			CC >PC
V-profiler			CC >PC	CC <PC	CC <PC	CC >PC
V-satellite	CC >PC		CC >PC	CC >PC	CC >PC	CC <PC
T-Aircraft	CC >PC	CC >PC	CC >PC		CC >PC	CC >PC
T-Rawinsonde	CC <PC		CC >PC	CC <PC		CC >PC

FIG. 8: Statistics of each distribution of observation impact for each DA system, evaluation forecast time, variable, and source. Mean, median and standard deviation (STD) were calculated from data trimmed between the 99th and 1st percentiles. The highlighted cells represent statistically significant differences between the statistics of CC and PC. CC >PC cells filled with dark grey while CC <PC cells filled with light grey. For the mean and median, smaller values indicate a greater positive observation impact. The mean was tested with the t-student test, the median with the MWW rank-sum test, and STD with Levene's test. All tests were computed with a significance level of 95%.

of the PC and CC DA systems' observation impact distributions were statistically significant. For the means, the t-student (Wilks 2011) test was used; for the medians, once again the MWW test was employed; and for the standard deviation, Levene's (Levene 1960) test was used. All tests were computed with a significance level of 95%. Given that U, V, and T were most numerous and had the largest observation impacts (e.g., Fig. 6), we focus on these three variables.

For 1-hour and 6-hour evaluation forecast times the observation impact distributions for the CC EnKF have longer tails than those of the PC EnKF, and accordingly, standard deviations of the observation impact are statistically significantly larger in the CC EnKF than in the PC EnKF for all variables and sources (Figs. 7,8). This suggests that CC forecasts have a greater change in the forecast error and greater errors (forecast - RAP), so the forecasts are further away from the RAP analysis compared to PC forecasts. These results align with those analyzed in Figures 3 and 4. Moreover, the CC EnKF distributions are typically left-skewed, indicating that more observations have a very large beneficial impact than a very large detrimental impact (Fig. 7).

The greater variability in the CC EnKF is likely a reflection of its analyses being less constrained than PC EnKF analyses, as PC analyses at 0000 UTC are still impacted by global analyses that

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-24-0127.1.

assimilated many more observations than either limited-area EnKF (S22). Less constraint suggests there was more variability in the CC EnKF state than in the PC EnKF state, which was manifested by more varied observation impacts in the CC EnKF relative to the PC EnKF.

Both median and mean values of the U, V, and T distributions (Fig. 7) are negative: between -0.0155 JKg^{-1} and -0.0012 JKg^{-1} for median values, between -50.0757 JKg^{-1} and -9.3763 JKg^{-1} for mean values. Overall, the assimilation of these variables reduces the forecast errors on both evaluation forecast times and DA systems, as evidenced by the mean and median values being negative. U wind component observations for CC have a statistically significantly more beneficial effect than for PC (more negative). Accordingly, when these observations are assimilated, the change in the forecast error for CC is reduced more than for PC. V wind component observations are not statistically significantly different between PC and CC. Conversely, for T observations, the median of PC is more beneficial than CC (more negative/smaller), meaning that assimilating these observations reduces the forecast error for PC to a greater extent than for CC.

To identify which observation sources of these three variables provide the most beneficial impact, the same statistics were computed separately for aircraft, rawinsondes, profilers, and satellite-tracked winds (Fig. 8).

In examining the mean and median, our analysis reveals mixed results. There are no statistically significant disparities between CC and PC for the medians of the EFSOI results pertaining to U and V wind components for the 1-hour evaluation forecast time (Fig. 8). However, when evaluating the 6-hour evaluation forecast time, there are some statistically significant differences regarding wind. For instance, in the case of V-profiler observations, CC has a statistically significantly smaller median observation impact than PC, and vice versa for V-satellite observations.

Overall, analysis of the observation impact distributions did not indicate that either DA system systematically used observations better than the other. This consistently suggests that differences in observation impact are not strongly influenced by the assimilation process itself. Besides, the CC background is generally less accurate than the PC background, allowing observations to potentially have a larger relative impact in CC; the EFSOI analysis may not capture this subtle difference when using cumulative observation impact.

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-24-0127.1.

4. Summary and conclusions

In this study, the EFSOI method introduced by Kalnay et al. (2012) was applied to PC and CC EnKF DA systems, each with 80 members and 15-km horizontal grid spacing, over a computational domain spanning the entire CONUS. We evaluated the observation impacts on forecasts at 1-hour, 6-hour and 12-hour lead times across 23 forecasts to explore whether the two DA systems utilize observations differently.

Performance metrics relative to RAP analyses (RMSE, bias, and MTE) indicate that PC forecasts generally outperformed CC forecasts. However, the EFSOI analysis showed that observation impacts were beneficial for both systems at all lead times, with differences between PC and CC decreasing as the forecast lead time increased. These results suggest that initial differences in observation impact may stem from PC's closer alignment with global ensemble background information in earlier cycles.

Further examination of observation sources revealed that PC and CC performed similarly for the most impactful subsets, such as rawinsondes and aircraft, where differences were not statistically significant. In contrast, statistically significant differences appeared for smaller observation sources with limited impact, like mesonet observations. This highlights that PC and CC assimilated critical observation sources comparably, while differences in less impactful observations were inconsistent. These findings emphasize the complexity of observation impacts and the need for ongoing optimization of observational strategies for better forecasts.

The assimilation of variables such as U, V, T, and RH showed an overall forecast improvement across both systems, with significant differences between CC and PC across the forecast times. However, no consistent signal emerged to explain the forecast differences between PC and CC. Upper-air observations were consistently beneficial, while surface observations tended to have detrimental impacts, likely due to localized noise and errors.

Finally, our analysis introduces further insights through the examination of EFSOI statistics distributions of U,V and T. They revealed greater variability in observation impacts for CC, suggesting less constraint in its analyses. While mean and median EFSOI statistics favored CC for U and V wind components, T observations favored PC. These asymmetrical distributions highlight differences in how each system assimilated certain observations, though these differences did not translate into systematic superiority for either system.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-24-0127.1.

Crucially, these results underscore that the differences in forecast performance between PC and CC were likely driven by factors beyond observation use. Broader DA system processes, such as the ingestion of global fields in the PC system and their influence on the large-scale model state, appear to play a key role. This influence principally includes a more accurate initialization of the large-scale model state, which interacts with the model's dynamical and physical processes to yield better forecasts. Additionally, the periodic re-initialization in the PC system helps prevent error propagation in forecasts. These factors are not fully captured by EFSOI statistics, complicating the interpretation of differences in observation impacts.

Operational services often use either PC or CC systems, making it critical to understand how their fundamental differences influence forecast outcomes. While EFSOI offers a useful diagnostic tool, its limitations in capturing some system processes, as mentioned before, suggest that further research is needed to isolate the sources of performance differences. The findings in this manuscript may not extend directly to other PC and CC systems with different cycle lengths. Even when considering the same PC and CC systems, if the evaluation is taken into account for forecasts initialized from analyses of the earlier cycles in PC, the observation impacts may differ from the ones obtained for forecasts initialized from the last cycling analyses in PC.

EFSOI remains valuable as an initial diagnostic for exploring observation impacts and guiding future experiments. However, to enhance the effectiveness of DA systems, assimilation techniques for certain observation sources and variables should be carefully reviewed and tailored to specific forecast objectives and operational contexts.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-24-0127.1.

Acknowledgments. This work was carried out within the framework of the Gimena Casaretto Doctoral Scholarship funded by CONICET and partially funded by Proyecto de Investigación Científica y Tecnológico de Aplicación Intensiva PICT-2021-CAT-I-00130. NSF NCAR is sponsored by the United States National Science Foundation. This work was partially funded by NSF NCAR's Short-term Explicit Prediction (STEP) program. The data assimilation systems and forecasts were executed on NSF NCAR's Cheyenne supercomputer (Computational and Information Systems Laboratory 2017). Computational and Information Systems Laboratory, 2017: Cheyenne: HPE/SGI ICE XA System (NCAR Community Computing). Boulder, CO: National Center for Atmospheric Research. https://doi.org/10.5065/D6RX99HX.

References

Anderson, J. L., T. Hoar, K. Raeder, H. Liu, N. Collins, R. Torn, and A. Arellano, 2009: The data assimilation research testbed: A community facility. *Bull. Amer. Meteor. Soc.*, **90**, 1283–1296.

Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55, https://doi.org/https://doi.org/10.1038/nature14956.

Benjamin, S. G., and Coauthors, 2016: A north american hourly assimilation and model forecast cycle: The rapid refresh. *Monthly Weather Review*, **144 (4)**, 1669 – 1694, https://doi.org/10.1175/MWR-D-15-0242.1.

Cardinali, C., 2018: Forecast sensitivity observation impact with an observation-only based objective function. *Quarterly Journal of the Royal Meteorological Society*, **144 (716)**, 2089–2098, https://doi.org/https://doi.org/10.1002/qj.3305, https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3305.

Casaretto, G., M. E. Dillon, Y. García Skabar, J. J. Ruiz, and M. Sacco, 2023: Ensemble forecast sensitivity to observations impact (efsoi) applied to a regional data assimilation system over south-eastern south america. *Atmospheric Research*, **295**, 106 996, https://doi.org/https://doi.org/10.1016/j.atmosres.2023.106996.

Cavallo, S. M., J. Berner, and C. Snyder, 2016: Diagnosing model errors from time-averaged tendencies in the weather research and forecasting (wrf) model. *Monthly Weather Review*, **144 (2)**, 759 – 779, https://doi.org/10.1175/MWR-D-15-0120.1.

Chen, T.-C., and E. Kalnay, 2020: Proactive quality control: Observing system experiments using the ncep global forecast system. *Monthly Weather Review*, **148 (9)**, 3911 – 3931, https://doi.org/ https://doi.org/10.1175/MWR-D-20-0001.1.

Dowell, D. C., and Coauthors, 2022: The high-resolution rapid refresh (hrrr): An hourly updating convection-allowing forecast model. part i: Motivation and system description. *Weather and Forecasting*, **37 (8)**, 1371 – 1395, https://doi.org/10.1175/WAF-D-21-0151.1.

Ehrendorfer, M., R. M. Errico, and K. D. Raeder, 1999: Singular-vector perturbation growth in a primitive equation model with moist physics. *Journal of the Atmospheric Sciences*, **56 (11)**, 1627 – 1648, https://doi.org/10.1175/1520-0469(1999)056⟨1627:SVPGIA⟩2.0.CO;2.

Gustafsson, N., and Coauthors, 2018: Survey of data assimilation methods for convective-scale numerical weather prediction at operational centres. *Quarterly Journal of the Royal Meteorological Society*, **144 (713)**, 1218–1256, https://doi.org/https://doi.org/10.1002/qj.3179.

Hersbach, H., and Coauthors, 2020: The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, **146 (730)**, 1999–2049, https://doi.org/https://doi.org/10.1002/qj.3803, https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3803.

Hotta, D., T.-C. Chen, E. Kalnay, Y. Ota, and T. Miyoshi, 2017: Proactive qc: A fully flow-dependent quality control scheme based on efso. *Monthly Weather Review*, **145 (8)**, 3331 – 3354, https://doi.org/10.1175/MWR-D-16-0290.1.

Houtekamer, P. L., and F. Zhang, 2016: Review of the ensemble kalman filter for atmospheric data assimilation. *Monthly Weather Review*, **144 (12)**, 4489 – 4532, https://doi.org/10.1175/ MWR-D-15-0440.1.

Hsiao, L.-F., D.-S. Chen, Y.-H. Kuo, Y.-R. Guo, T.-C. Yeh, J.-S. Hong, C.-T. Fong, and C.-S. Lee, 2012: Application of wrf 3dvar to operational typhoon prediction in taiwan: Impact of outer loop and partial cycling approaches. *Weather and Forecasting*, **27 (5)**, 1249 – 1263, https://doi.org/10.1175/WAF-D-11-00131.1.

Hu, G., S. L. Dance, R. N. Bannister, H. G. Chipilski, O. Guillet, B. Macpherson, M. Weissmann, and N. Yussouf, 2023: Progress, challenges, and future steps in data assimilation for convection-permitting numerical weather prediction: Report on the virtual meeting held on 10 and 12

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-24-0127.1.

november 2021. *Atmospheric Science Letters*, **24 (1)**, e1130, https://doi.org/https://doi.org/10.1002/asl.1130, https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/asl.1130.

James, E. P., and S. G. Benjamin, 2017: Observation system experiments with the hourly updating rapid refresh model using gsi hybrid ensemble–variational data assimilation. *Monthly Weather Review*, **145 (8)**, 2897 – 2918, https://doi.org/10.1175/MWR-D-16-0398.1.

Kalnay, E., Y. Ota, T. Miyoshi, and J. Liu, 2012: A simpler formulation of forecast sensitivity to observations: application to ensemble kalman filters. *Tellus A: Dynamic Meteorology and Oceanography*, **64 (1)**, 18 462, https://doi.org/10.3402/tellusa.v64i0.18462.

Kotsuki, S., K. Kurosawa, and T. Miyoshi, 2019: On the properties of ensemble forecast sensitivity to observations. *Quarterly Journal of the Royal Meteorological Society*, **145 (722)**, 1897–1914, https://doi.org/https://doi.org/10.1002/qj.3534, https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3534.

Kunii, M., T. Miyoshi, and E. Kalnay, 2012: Estimating the impact of real observations in regional numerical weather prediction using an ensemble kalman filter. *Monthly Weather Review*, **140 (6)**, 1975 – 1987, https://doi.org/10.1175/MWR-D-11-00205.1.

Levene, H., 1960: Robust tests for equality of variances. *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, I. Olkin, Ed., Stanford University Press, 278–292.

Lien, G.-Y., D. Hotta, E. Kalnay, T. Miyoshi, and T.-C. Chen, 2018: Accelerating assimilation development for new observing systems using efso. *Nonlinear Processes in Geophysics*, **25 (1)**, 129–143, https://doi.org/10.5194/npg-25-129-2018.

Lorenc, A. C., and R. T. Marriott, 2014: Forecast sensitivity to observations in the met office global numerical weather prediction system. *Quarterly Journal of the Royal Meteorological Society*, **140 (678)**, 209–224, https://doi.org/https://doi.org/10.1002/qj.2122.

Ota, Y., J. C. Derber, E. Kalnay, and T. Miyoshi, 2013: Ensemble-based observation impact estimates using the ncep gfs. *Tellus A: Dynamic Meteorology and Oceanography*, **65 (1)**, 20 038, https://doi.org/10.3402/tellusa.v65i0.20038.

Poterjoy, J., G. J. Alaka, and H. R. Winterbottom, 2021: The irreplaceable utility of sequential data assimilation for numerical weather prediction system development: Lessons learned from

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-24-0127.1.

an experimental hwrf system. *Weather and Forecasting*, **36 (2)**, 661 – 677, https://doi.org/10.1175/WAF-D-20-0204.1.

Powers, J. G., 2017: The weather research and forecasting model: Overview, system efforts, and future directions. *Bull. Amer. Meteor. Soc.*, **98**, 1717–1737.

Romine, G. S., C. S. Schwartz, C. Snyder, J. L. Anderson, and M. L. Weisman, 2013: Model bias in a continuously cycled assimilation system and its influence on convection-permitting forecasts. *Monthly Weather Review*, **141 (4)**, 1263 – 1284, https://doi.org/10.1175/MWR-D-12-00112.1.

Schraff, C., H. Reich, A. Rhodin, A. Schomburg, K. Stephan, A. Periáñez, and R. Potthast, 2016: Kilometre-scale ensemble data assimilation for the cosmo model (kenda). *Quarterly Journal of the Royal Meteorological Society*, **142 (696)**, 1453–1472, https://doi.org/https://doi.org/10.1002/qj.2748, https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.2748.

Schwartz, C. S., J. Poterjoy, J. R. Carley, D. C. Dowell, G. S. Romine, and K. Ide, 2022: Comparing partial and continuously cycling ensemble kalman filter data assimilation systems for convection-allowing ensemble forecast initialization. *Weather and Forecasting*, **37 (1)**, 85 – 112, https://doi.org/10.1175/WAF-D-21-0069.1.

Skamarock, W., J. Klemp, J. Dudhia, D. Gill, D. Barker, W. Wang, and J. Powers, 2008: A description of the advanced research WRF version 3. **27**, 3–27.

Sommer, M., and M. Weissmann, 2014: Observation impact in a convective-scale localized ensemble transform kalman filter. *Quarterly Journal of the Royal Meteorological Society*, **140 (685)**, 2672–2679, https://doi.org/https://doi.org/10.1002/qj.2343.

Torn, R. D., and C. A. Davis, 2012: The influence of shallow convection on tropical cyclone track forecasts. *Monthly Weather Review*, **140 (7)**, 2188 – 2197, https://doi.org/10.1175/MWR-D-11-00246.1.

Wilks, D. S., 2011: *Statistical methods in the atmospheric sciences*. Academic press.

Yamazaki, A., T. Miyoshi, J. Inoue, T. Enomoto, and N. Komori, 2021: Efso at different geographical locations verified with observing-system experiments. *Weather and Forecasting*, https://doi.org/10.1175/WAF-D-20-0152.1.