Proactive QC: A Fully Flow-Dependent Quality Control Scheme Based on EFSO

DAISUKE HOTTA

University of Maryland, College Park, College Park, Maryland, and Japan Meteorological Agency, Tokyo, Japan

TSE-CHUN CHEN AND EUGENIA KALNAY

University of Maryland, College Park, College Park, Maryland

YOICHIRO OTA

Japan Meteorological Agency, Tokyo, Japan

TAKEMASA MIYOSHI

University of Maryland, College Park, College Park, Maryland, and RIKEN Advanced Institute for Computational Science, Kobe, Japan

(Manuscript received 29 July 2016, in final form 3 April 2017)

ABSTRACT

Despite dramatic improvements over the last decades, operational NWP forecasts still occasionally suffer from abrupt drops in their forecast skill. Such forecast skill "dropouts" may occur even in a perfect NWP system because of the stochastic nature of NWP but can also result from flaws in the NWP system. Recent studies have shown that dropouts occur due not to a model's deficiencies but to misspecified initial conditions, suggesting that they could be mitigated by improving the quality control (QC) system so that the observation-minus-background (O-B) innovations that would degrade a forecast can be detected and rejected. The ensemble forecast sensitivity to observations (EFSO) technique enables for the quantification of how much each observation has improved or degraded the forecast. A recent study has shown that 24-h EFSO can detect detrimental O-B innovations that caused regional forecast skill dropouts and that the forecast can be improved by not assimilating them. Inspired by that success, a new QC method is proposed, termed proactive QC (PQC), that detects detrimental innovations 6 h after the analysis using EFSO and then repeats the analysis and forecast without using them. PQC is implemented and tested on a lower-resolution version of NCEP's operational global NWP system. It is shown that EFSO is insensitive to the choice of verification and lead time (24 or 6 h) and that PQC likely improves the analysis, as attested to by forecast improvements of up to 5 days and beyond. Strategies for reducing the computational costs and further optimizing the observation rejection criteria are also discussed.

1. Introduction

Numerical weather prediction (NWP) has gone through dramatic improvement over the last several decades (e.g., Simmons 2011). Despite the very high average forecast skill, however, current operational NWP systems still suffer from abrupt drops in forecast performance (e.g., Alpert et al. 2009; Kumar et al. 2009; Rodwell et al. 2013). Such forecast skill *dropouts*, or *busts*, are highly undesirable because they not only degrade the average forecast skills but also taint the operational reliability of the NWP forecasts.

There are two contrasting views on why forecast dropouts occur. One, purely probabilistic interpretation views a deterministic analysis as just a single stochastic draw from the hypothetical population whose probability distribution function obeys that of a minimum-variance or maximum-likelihood estimator conditioned by the particular realizations of the first guess and observations given at the initial time; from this perspective, dropouts

Denotes content that is immediately available upon publication as open access.

Corresponding author: Daisuke Hotta, dhotta@mri-jma.go.jp

DOI: 10.1175/MWR-D-16-0290.1

may occur, even with an optimal data assimilation (DA) system and flawless observations, if the analysis increment happens to project strongly on rapidly growing modes of the analysis errors in such a way as to amplify them [Lorenc and Marriott (2014, section 5); see also the discussion in the appendix below]. Another interpretation acknowledges the imperfection of the operational system and tries to attribute dropouts to particular flaws in the system. In this line, recent studies by the National Centers for Environmental Prediction (NCEP) "Dropout Team" (Alpert et al. 2009; Kumar et al. 2009) have shown that many dropouts tend to occur due not to model deficiencies but to the misspecification of the initial conditions and they argued that one promising way to alleviate the dropout problem is to improve the operational quality control (QC) system. We believe that the two aspects are both important in understanding dropouts. In interpreting the results shown later in this paper, we will pay attention to both possibilities.

Advances in QC methods have improved the accuracy of NWP (e.g., Kalnay 2003, section 5.8) but current operational QC methods still have room for improvement. The biggest limitation is that the current operational approach, which first screens out the observations whose observation-minus-background (O-B) innovations exceed predetermined thresholds ("gross-error check") and then assigns smaller weights to the observations that are inconsistent with the (tentative) analysis ("nonlinear QC"; see the last paragraph of section 2a), can mistakenly screen out accurate observations and allow inaccurate observations to be used in "latent dropout" situations, where the background is unreliable and the other observations in the vicinity are either inaccurate or not available. Flow-dependent techniques such as dynamic QC, which is discussed by Onogi (1998), which allows the thresholds to vary depending on the estimated background accuracy, can alleviate this issue but only partially. A fully flow-dependent QC method that filters out only the observations that significantly degrade forecasts is thus needed, but such a method requires knowing in advance whether an observation will improve or degrade a forecast.

Langland and Baker (2004, hereafter LB04) made a breakthrough in this direction by introducing a diagnostic method, called forecast sensitivity to observations (FSO), which enables to estimate, at a computationally feasible cost, how much each observation improved or degraded the 24-h forecast. Their formulation exploits an adjoint sensitivity technique and is applicable to variational DA systems. Major operational NWP centers soon adopted this technique (e.g., Cardinali 2009; Gelaro and Zhu 2009; Ishibashi 2010; Lorenc and Marriott 2014) and showed that it is a powerful diagnostic. Its ensemble-based formulation, ensemble FSO (EFSO), was devised by Liu and Kalnay (2008) and Li et al. (2010) for the local ensemble transform Kalman filter (LETKF; Hunt et al. 2007); Kalnay et al. (2012) derived a simpler and more accurate new EFSO formulation applicable not only to LETKF but to any ensemble Kalman filter (EnKF). Ota et al. (2013, hereafter ODKM13) successfully implemented the new EFSO on NCEP's quasi-operational global EnKF system and showed that EFSO is consistent with previous adjoint-based FSO studies. Furthermore, they applied EFSO to individual cases and succeeded in attributing regional forecast dropouts to specific O-B innovations. Strikingly, in one of their retrospective data-denial experiments based on 24-h EFSO, the regional 24-h forecast error was reduced by as much as 30% by not assimilating the observations that showed large negative EFSO impacts.

We emphasize here that we should not interpret a large negative (E)FSO impact from an observation as a direct indication of any flaws in that observation. Unlike what its name may suggest, (E)FSO estimates the impact on a forecast from the *O-B innovation* ($\delta\overline{\mathbf{y}}_0^{\mathrm{ob}}$) associated with each observation rather than the observation itself; detrimental impacts can thus arise from both an erroneous background and erroneous observations. Moreover, because of the stochastic nature of DA, even a perfectly benign observation and background within a perfect DA system could show detrimental impact, and the impact can be large regardless of the actual errors of the observation or background if the observation is made in a region with high sensitivity. Furthermore, detrimental impact can be caused by many reasons[1] other than issues in the observations themselves. In this manuscript, we refer to the observations that, if assimilated, would result in significant forecast degradation by the expression " (observations associated with) detrimental $\delta\overline{\mathbf{y}}_0^{\mathrm{ob}}$ innovations" to avoid giving a negative connotation to such observations themselves.

Following the success of ODKM13, in this study we propose a simple, new QC scheme which we denote "proactive QC" (PQC), which exploits EFSO's capacity to identify detrimental $\delta\overline{\mathbf{y}}_0^{\mathrm{ob}}$ innovations that actually

[1] These include (i) large measurement errors of the instrument; (ii) imperfection in the retrieval algorithm; (iii) errors of the observation operator, including representativeness errors; (iv) forecast model errors; (v) suboptimal preprocessing, including thinning, bias correction, and gross-error check; and (vi) incorrect specification of the observation error covariance. Note that detrimental impacts may not be necessarily due to problems in the background or the observations (like in the first and second entries in the list above) but rather to problems in the DA system (like in the third through sixth issues).

degrade the forecast skill. We first perform DA using all the available observations that passed the standard QC and 6 h later we compute regional 6-h forecast errors (with respect to the analysis) and apply an algorithm to detect regional skill dropouts. We next conduct EFSO diagnostics on the detected regions to identify potential detrimental $\delta\overline{\mathbf{y}}_0^{\mathrm{ob}}$ innovations that are likely responsible for the regional dropouts. Finally, we repeat the analysis and 6-h forecast *without* assimilating the identified detrimental $\delta\overline{\mathbf{y}}_0^{\mathrm{ob}}$ innovations to provide a better first guess to the analysis at the next cycle.

To implement this PQC scheme in an operational system, we need to address several important questions:

1) Does EFSO work for an ensemble–variational hybrid DA system? Hybrid approaches have been adopted by several operational NWP centers (Buehner 2005; Kleist 2012; Wang et al. 2013; Kleist and Ide 2015a; Clayton et al. 2013) but EFSO has not yet been tested on such a system.
2) Is a short lead time of 6 h long enough to capture meaningful signals from detrimental $\delta\overline{\mathbf{y}}_0^{\mathrm{ob}}$ innovations, since analysis errors may not be negligible compared with 6-h forecast errors?
3) How do we choose the observations to reject given the EFSO impacts of individual observations?
4) Does rejection of the observations identified by EFSO as detrimental improve the analysis and the first guess in the next cycle?

This paper aims to show that PQC does improve the analysis in an operational system, providing answers to questions 1–4 by conducting experiments using a lower-resolution version of the NCEP's operational global NWP system. Section 2 reviews the EFSO algorithm following Kalnay et al. (2012) and describes the proposed algorithm of PQC. Section 3 describes the experimental settings. Section 4 shows the EFSO's dependence on verifying truth and evaluation lead time, providing answers to questions 1 and 2 above. Section 5 describes the data-denial experiments and addresses questions 3 and 4. Section 6 gives a summary and offers our conclusions, including several ideas on how to further reduce computational costs in an operational implementation.

2. EFSO formulation and PQC algorithm

a. EFSO formulation following Kalnay et al. (2012)

Denoting the gain matrix by $\mathbf{K}$, the analysis equation can be written as

$$\overline{\mathbf{x}}_0^a - \overline{\mathbf{x}}_0^b = \mathbf{K}\delta\overline{\mathbf{y}}_0^{\mathrm{ob}}, \tag{1}$$

where $\overline{\mathbf{x}}_0^a$ and $\overline{\mathbf{x}}_0^b$ are the ensemble mean analysis and background, respectively, and $\delta\overline{\mathbf{y}}_0^{\mathrm{ob}} = \mathbf{y}_0^o - \overline{H(\mathbf{x}_0^b)}$ is the O-B innovation of the ensemble mean, with $H(\cdot)$ denoting the observation operator, all valid at time $t = 0$. Unlike variational methods, EnKF allows us to directly estimate the gain matrix $\mathbf{K}$ by

$$\mathbf{K} = \mathbf{A}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1} \approx \frac{1}{K-1}(\mathbf{X}^a\mathbf{X}^{a\mathrm{T}})\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1} \approx \frac{1}{K-1}\mathbf{X}^a\mathbf{Y}^{a\mathrm{T}}\mathbf{R}^{-1}, \tag{2}$$

where K is the ensemble size; $\mathbf{A}$ and $\mathbf{R}$ are the analysis and observation error covariance matrices, respectively; $\mathbf{H}$ is the Jacobian of H; $\mathbf{X}^a$ is the matrix of the analysis perturbations valid at time 0; and $\mathbf{Y}^a = \mathbf{H}\mathbf{X}^a$. Using (2), the analysis equation (1) yields

$$\overline{\mathbf{x}}_0^a - \overline{\mathbf{x}}_0^b \approx \frac{1}{K-1}\mathbf{X}^a\mathbf{Y}^{a\mathrm{T}}\mathbf{R}^{-1}\delta\overline{\mathbf{y}}_0^{\mathrm{ob}}. \tag{3}$$

Now, following LB04, we measure the change of the forecast error due to the assimilation by

$$\begin{aligned}\Delta\mathbf{e}^2 &= \mathbf{e}_{t|0}^{\mathrm{T}}\mathbf{C}\mathbf{e}_{t|0} - \mathbf{e}_{t|-6}^{\mathrm{T}}\mathbf{C}\mathbf{e}_{t|-6}\\ &= (\mathbf{e}_{t|0} - \mathbf{e}_{t|-6})^{\mathrm{T}}\mathbf{C}(\mathbf{e}_{t|0} + \mathbf{e}_{t|-6}),\end{aligned} \tag{4}$$

with

$$\mathbf{e}_{t|0} = \overline{\mathbf{x}}_{t|0} - \mathbf{x}_t^v, \quad \text{and} \quad \mathbf{e}_{t|-6} = \overline{\mathbf{x}}_{t|-6} - \mathbf{x}_t^v, \tag{5}$$

where $\overline{\mathbf{x}}_{t|-6}$ and $\overline{\mathbf{x}}_{t|0}$ denote the ensemble mean forecast valid at time t initialized, respectively, at time -6 and 0 (i.e., before and after the assimilation), $\mathbf{x}_t^v$ denotes the verifying state at time t, and $\mathbf{C}$ is a square matrix that defines the error norm (section 3). Denoting the forecast operator that advances the model state from t_1 to t_2 by $M_{t_2|t_1}(\cdot)$ and its Jacobian by $\mathbf{M}_{t_2|t_1}$, and using (3), we have

$$\begin{aligned}\Delta\mathbf{e}^2 &= (\overline{\mathbf{x}}_{t|0} - \overline{\mathbf{x}}_{t|-6})^{\mathrm{T}}\mathbf{C}(\mathbf{e}_{t|0} + \mathbf{e}_{t|-6}) \approx [\mathbf{M}_{t|0}(\overline{\mathbf{x}}_0^a - \overline{\mathbf{x}}_0^b)]^{\mathrm{T}}\mathbf{C}(\mathbf{e}_{t|0} + \mathbf{e}_{t|\text{-}6})\\ &\approx \frac{1}{K-1}(\mathbf{M}_{t|0}\mathbf{X}^a\mathbf{Y}^{a\mathrm{T}}\mathbf{R}^{-1}\delta\overline{\mathbf{y}}_0^{\mathrm{ob}})^{\mathrm{T}}\mathbf{C}(\mathbf{e}_{t|0} + \mathbf{e}_{t|-6}) \approx \delta\overline{\mathbf{y}}_0^{\mathrm{ob}\mathrm{T}}\frac{1}{K-1}\mathbf{R}^{-1}\mathbf{Y}^a\mathbf{X}_{t|0}^{f\mathrm{T}}\mathbf{C}(\mathbf{e}_{\mathrm{t}|0} + \mathbf{e}_{t|-6}),\end{aligned} \tag{6}$$

where $\mathbf{X}_{t|0}^f$ is the matrix of forecast perturbations initialized at time 0 and valid at time t. Equation (6) can be interpreted as an inner product of the O-B innovation vector $\delta\overline{\mathbf{y}}_0^{\mathrm{ob}}$ and a sensitivity vector:

$$\Delta\mathbf{e}^2 \approx \delta\overline{\mathbf{y}}_0^{\mathrm{ob}\mathrm{T}}\frac{\partial(\Delta\mathbf{e}^2)}{\partial\mathbf{y}}, \tag{7}$$

where

$$\frac{\partial(\Delta\mathbf{e}^2)}{\partial\mathbf{y}} = \frac{1}{K-1}\mathbf{R}^{-1}\mathbf{Y}^{a}\mathbf{X}_{t|0}^{f\mathrm{T}}\mathbf{C}(\mathbf{e}_{t|0}+\mathbf{e}_{t|-6}). \tag{8}$$

Thus, the change in the forecast error due to the assimilation of observations $\mathbf{y}_0^o$ can be decomposed into a sum of contributions from the O-B innovations associated with each observation. The contribution from a single observation $y_{0,l}^o$, the lth element of $\mathbf{y}_0^o$, or more precisely, the contribution from its corresponding O-B innovation $\delta\overline{y}_{0,l}^{\mathrm{ob}}$, is

$$(\Delta\mathbf{e}^2)\big|_{y_{0,l}^o} \approx \delta\overline{y}_{0,l}^{\mathrm{ob}}\frac{\partial(\Delta\mathbf{e}^2)}{\partial y_l}. \tag{9}$$

This is the EFSO impact that we wish to employ.

As with any EnKF, localization needs to be applied to the error covariances to suppress sampling errors whenever K is smaller than the number of degrees of freedom of the predicted dynamical system. After applying localization, the sensitivity vector becomes

$$\frac{\partial(\Delta\mathbf{e}^2)}{\partial\mathbf{y}} = \frac{1}{K-1}\mathbf{R}^{-1}[\boldsymbol{\rho}\circ(\mathbf{Y}^{a}\mathbf{X}_{t|0}^{f\mathrm{T}})]\mathbf{C}(\mathbf{e}_{t|0}+\mathbf{e}_{t|-6}), \tag{10}$$

where the symbol $\circ$ represents element-wise multiplication (Schur product) and $\boldsymbol{\rho}$ is a matrix whose (l, j) element is a localization factor of the lth observation onto the jth grid point.

Unlike in 4DVar, in EnKF, an explicit estimation of the analysis error covariance and the gain matrix are available. As pointed out by ODKM13, we can exploit this to approximately estimate how the analysis and forecast would change by not assimilating a given subset of the observations. Let $\delta\overline{\mathbf{y}}_0^{\mathrm{ob,deny}}$ be a column vector whose elements corresponding to the denied observations are identical to those of $\delta\overline{\mathbf{y}}_0^{\mathrm{ob}}$ but others are all set to zero. Then, assuming that the Kalman gain $\mathbf{K}$ does not change much by excluding the denied observations (which should be valid if the fraction of denied observations is small; see section 6 for a discussion), the analysis that would be obtained by not assimilating them can be approximated by

$$\overline{\mathbf{x}}_0^{a,\mathrm{deny}} \approx \overline{\mathbf{x}}_0^{b} + \mathbf{K}(\delta\overline{\mathbf{y}}_0^{\mathrm{ob}} - \delta\overline{\mathbf{y}}_0^{\mathrm{ob,deny}}) = \overline{\mathbf{x}}_0^{a} - \mathbf{K}\delta\overline{\mathbf{y}}_0^{\mathrm{ob,deny}}. \tag{11}$$

Substituting (2) and applying localization as in (10), we obtain

$$\overline{\mathbf{x}}_0^{a,\mathrm{deny}} - \overline{\mathbf{x}}_0^{a} \approx -\frac{1}{K-1}[\boldsymbol{\rho}\circ\mathbf{X}^{a}\mathbf{Y}^{a\mathrm{T}}]\mathbf{R}^{-1}\delta\overline{\mathbf{y}}_0^{\mathrm{ob,deny}}. \tag{12}$$

Similarly, the change in *forecast* can be approximated by

$$\overline{\mathbf{x}}_{t|0}^{f,\mathrm{deny}} - \overline{\mathbf{x}}_{t|0}^{f} \approx -\frac{1}{K-1}[\boldsymbol{\rho}\circ\mathbf{X}_{t|0}^{f}\mathbf{Y}^{a\mathrm{T}}]\mathbf{R}^{-1}\delta\overline{\mathbf{y}}_0^{\mathrm{ob,deny}}. \tag{13}$$

Several NWP systems use QC to ensure consistency of an observation with respect to the analysis or the other observations assimilated in the same cycle (Lorenc 1981; Ingleby and Lorenc 1993; Anderson and Järvinen 1999; Tavolato and Isaksen 2015). Such a QC method, known as nonlinear QC or variational QC (Anderson and Järvinen 1999), is implemented in the variational part of NCEP's global DA system and the resulting flags are used in the EnKF part. We note here that 0-h EFSO (or analysis sensitivity to observations) can also be used to retrospectively check whether the inconsistent observations were effectively rejected. Verifying against the analysis, (4) yields

$$\Delta\mathbf{e}^2 = -(\mathbf{K}\delta\overline{\mathbf{y}}_0^{\mathrm{ob}})^{\mathrm{T}}\mathbf{C}\delta\overline{\mathbf{x}}^{a,b}, \tag{14}$$

from which we can deduce that the contribution to $\Delta\mathbf{e}^2$ from a single observation is negative (positive) if the partial analysis increment attributable to that observation is consistent (inconsistent) with the total analysis increment. Thus, we expect that 0-h EFSO should be mostly negative (or should at least not show a large positive value) provided that the nonlinear QC works well. Liu et al. (2009) proposed a similar idea based on self-sensitivity diagnostics; in fact, their Eq. (13), which "predicts" what the analysis should be if a particular observation was not assimilated, can be obtained by applying the observation operator to our equation (12). We show an example of this diagnostics in section 4b, where we discuss EFSO's dependence on evaluation lead time.

b. PQC algorithm

PQC is based on the following idea: if the assimilation of some $\delta\overline{\mathbf{y}}_0^{\mathrm{ob}}$ innovations significantly degrades the forecast, such innovations should be identifiable by EFSO; we can then improve the analysis and forecast by not assimilating them. Let 0 h be the initial time for which PQC is to be applied and assume that the DA system has a 6-h assimilation window. The algorithm can be summarized as follows:

1) run the regular DA cycle from time −6 to 0 h and then from 0 to +6 h;
2) using the information available from step 1, detect horizontal regions where "forecast skill dropout" is likely to occur, using the empirical *regional dropout detection criteria* (section 2c);
3) if such regions are detected, perform 6-h EFSO targeting each of those regions;
4) then, apply *detrimental innovation selection criteria* (section 5b) to identify the observations whose $\delta\overline{\mathbf{y}}_0^{\mathrm{ob}}$ innovations are likely detrimental; and

TABLE 1. Experimental settings of our PQC experiments compared with ODKM13.

	This study	ODKM13
Forecast		
Forecast model	GFS	GFS
Resolution (deterministic)	T254L64	N/A
Resolution (ensemble)	T126L64	T254L64
Analysis		
DA system	GSI hybrid 3DVar with ensemble from LETKF	Pure EnSRF
No. of members	80	80
Assimilated observations	Same as the operational system	Same as the operational system but without precipitation retrieval from TRMM/TMI
Localization cutoff length	2000 km (horizontal) Twice the scale height (vertical)	2000 km (horizontal) Twice the scale height (vertical)
EFSO		
Verifying truth	GSI analysis or LETKF mean analysis	EnSRF mean analysis
Evaluation lead time	6, 12, and 24 h	24 h
Localization cutoff length	Same as LETKF	Same as EnSRF
Error norm	Dry and moist total energy	Dry and moist total energy
Period		
Spinup	7 days from 0000 UTC 1 Jan to 1800 UTC 7 Jan	7 days from 0000 UTC 1 Jan to 1800 UTC 7 Jan
Statistical verification	31 days from 0000 UTC 8 Jan to 1800 UTC 7 Feb	31 days from 0000 UTC 8 Jan to 1800 UTC 7 Feb
Case studies	34 days from 0000 UTC 8 Jan to 1800 UTC 10 Feb	31 days from 0000 UTC 8 Jan to 1800 UTC 7 Feb

5) if such detrimental $\delta\overline{\mathbf{y}}_0^{\mathrm{ob}}$ innovations are identified, repeat DA for time 0 h without using them [or use the computationally faster approximation (12)].

c. Regional dropout detection criteria

Within an operational context it would be computationally unfeasible to perform EFSO targeted at every region of the globe. We thus need to detect potential *regional forecast dropouts* without actually performing EFSO. We follow ODKM13's procedure to detect potential regional dropouts:

- divide the globe into 30° × 30° latitude–longitude rectangular cells allowing for overlaps by shifting the longitude by 10° and the latitude by 15°;
- for each of the cells, compute regional forecast errors for 6- and 12-h forecasts valid at the same time ($\mathbf{e}_{t|0}^{\mathrm{T}}\mathbf{C}\mathbf{e}_{t|0} =: e_{t|0}^{f}$ and $\mathbf{e}_{t|-6}^{\mathrm{T}}\mathbf{C}\mathbf{e}_{t|-6} =: e_{t|-6}^{f}$, where $t = 6$ h);
- find cells where both $e_{t|0}^{f}/\langle e_{t|0}^{f}\rangle$ (where the bracket $\langle\cdot\rangle$ represents the mean over the period of experimentation) and $e_{t|0}^{f}/e_{t|-6}^{f}$ are more than twice their standard deviations, where the standard deviation is evaluated over the period of experimentation; and
- if two or more overlapping or adjacent cells satisfy the above criteria, merge them to form a single dropout region.

Using this procedure, we were able to narrow down the regions needed to compute EFSO to only one or two regions per cycle while identifying more than 200 cases of potential regional dropouts.

3. Experimental setup

This section describes the setup of our experimentation. Since our experiments are an extension of ODKM13, we compare our setup with theirs in Table 1.

a. Forecast model and DA system

The forecast model we used is the NCEP GFS model that had been operational until January 2015. Because of limited computational resources, however, we use it with the reduced horizontal resolutions of T254 (~55 km) and T126 (~110 km), respectively, for deterministic and ensemble runs (as opposed to the operational T574 and T254 resolutions). As in the operational system, the analysis is produced by the two-way nested EnKF–3DVar hybrid Gridpoint Statistical Interpolation analysis system (GSI) (Wang et al. 2013; Kleist 2012; Kleist and Ide 2015a,b). For the ensemble generation, we adopt the LETKF instead of the operational serial ensemble square root filter (EnSRF) of Whitaker and Hamill (2002). We use an ensemble size of 80 and apply both localization and inflation to the covariance. The parameters for localization and inflation are identical to the operational serial EnSRF with T254L64 resolution: the covariance is localized using Gaspari and Cohn's (1999) function with a cutoff length of 2000 km for the horizontal and twice the scale height for the vertical (equivalent to *e*-folding scales of 800 km and 0.8 scale heights). For inflation, we adopt both "relaxation to prior spread" (RTPS) multiplicative inflation and a National Meteorological Center (NMC, now known as NCEP) method–like additive inflation, as described in Wang et al.

(2013), but with a relaxation parameter of 0.85 for RTPS and a scaling parameter of 0.32 for the additive inflation. Since the parameters we use in T126 LETKF are optimized for use with T254 serial EnSRF, these choices may not be optimal. Nevertheless, the system worked without any problem.

b. Observations

In this study, we assimilate all the observations that were assimilated in the operational system during the period of our experimentation. Comprehensive documentation of all the observation types assimilated in this study is available online (Keyser 2011, 2013). Conventional (i.e., nonradiance) data are grouped by "report types" (see Table 2 in Keyser 2013); satellite radiances are grouped by the sensors, as summarized in Table 1 in ODKM13.

c. EFSO configurations

We compute EFSO impacts for each assimilated observation using (9) and (10) for evaluation lead times of 0, 6, 12, and 24 h. We localize the cross covariance using the same localization function as in LETKF analysis and adopt the moving localization scheme of ODKM13. Since we are interested in EFSO over a shorter lead time (6 h as opposed to 24 h), the impact of localization advection should be much smaller.

It is a common practice in computing forecast error vectors with (5) to use the DA system's own analysis as the verifying truth, $\mathbf{x}_t^v$. An issue that arises when applying EFSO to a hybrid DA is that two different analyses are available, one from the variational part and the other from EnKF. In section 4c, we explore which to choose when performing EFSO. As the error metric $\mathbf{C}$, we adopt either the dry or moist total energy norm (Ehrendorfer et al. 1999); see also Eq. (9) in ODKM13. We perform DA cycles from 0000 UTC 1 January to 1800 UTC 10 February 2012 and discard the first 7 days to account for spinup. For the first cycle, we create the first guess by interpolating the operational (higher resolution) GSI product to our resolution. Note that this relatively short spinup period is sufficient in our case because the first guess ingested during the first cycle is already close to fully spun up. We perform statistical verifications for the 31-day period from 8 January to 8 February 2012, but for case studies presented in section 5 we include the whole 34-day period (excluding the spinup).

4. Sensitivity of EFSO to the choice of evaluation lead time and verifying truth

a. Consistency with previous FSO studies

Since this study is the first to apply EFSO to a hybrid DA system, it is important to make sure of the validity of EFSO within a hybrid DA framework. We first compare our EFSO results with those of ODKM13, whose experimental setup is similar to ours except for the resolution and for the use of a pure (nonhybrid) EnSRF. We then compare our results with an adjoint-based FSO, taking, as an example, the recent work at NASA/GMAO (Holdaway et al. 2014). As noted before, in the hybrid system we have two different choices for the verifying truth. We will show that the EFSO results do not depend significantly on this choice; thus, in this section, we concentrate on the EFSO impacts verified against the GSI analysis.

To compare with the results of ODKM13, we start with Figs. 1c,f showing the 24-h EFSO impacts from each observation type measured, respectively, by the moist and dry total energy norm (shorter EFSO results are discussed in the next subsection). The impacts are evaluated for the whole globe and are averaged over the 31-day verification period defined in section 3. Note that observations with positive impacts have negative EFSO values because they decrease the forecast errors. By comparing these figures with Figs. 2 and 3 in ODKM13, we find that, despite several differences in the experimental setup including their use of hybrid DA, our results are mostly consistent. In particular, the following notable features pointed out by ODKM13 are also valid with our results:

- The Advanced Microwave Sounding Unit A (AMSU-A) contributes most positively followed by aircraft, radiosondes and the Infrared Atmospheric Sounding Interferometer (IASI).
- Ozone contributes slightly negatively.
- All satellite radiance observations [especially the Microwave Humidity Sounder (MHS)] and piloted balloon (PIBAL) exhibit smaller impacts with the dry norm than with the moist norm.

Our results are also consistent with the adjoint-based FSO computed by Holdaway et al. (2014) (see their Figs. 6 and 9): despite differences in the verification period, the sampling (6 h in ours and daily in theirs), and the definition of the error norm, our results and those of Holdaway et al. (2014) agree in that AMSU-A, aircraft, and radiosondes are the top three positively contributing types, followed by IASI and AIRS, and that the impacts from satellite radiances are ordered, from largest to smallest, as AMSU-A, IASI, AIRS, HIRS, and MHS. The consistency of our results with respect to ODKM13 and Holdaway et al. (2014) strongly supports the validity of using EFSO within a hybrid DA system.

b. Dependence on evaluation lead time

Our PQC algorithm relies on the ability of 6-h EFSO to detect detrimental $\delta\overline{\mathbf{y}}_0^{\mathrm{ob}}$ innovations. Thus, it is of vital importance to explore the characteristics of 6-h EFSO,

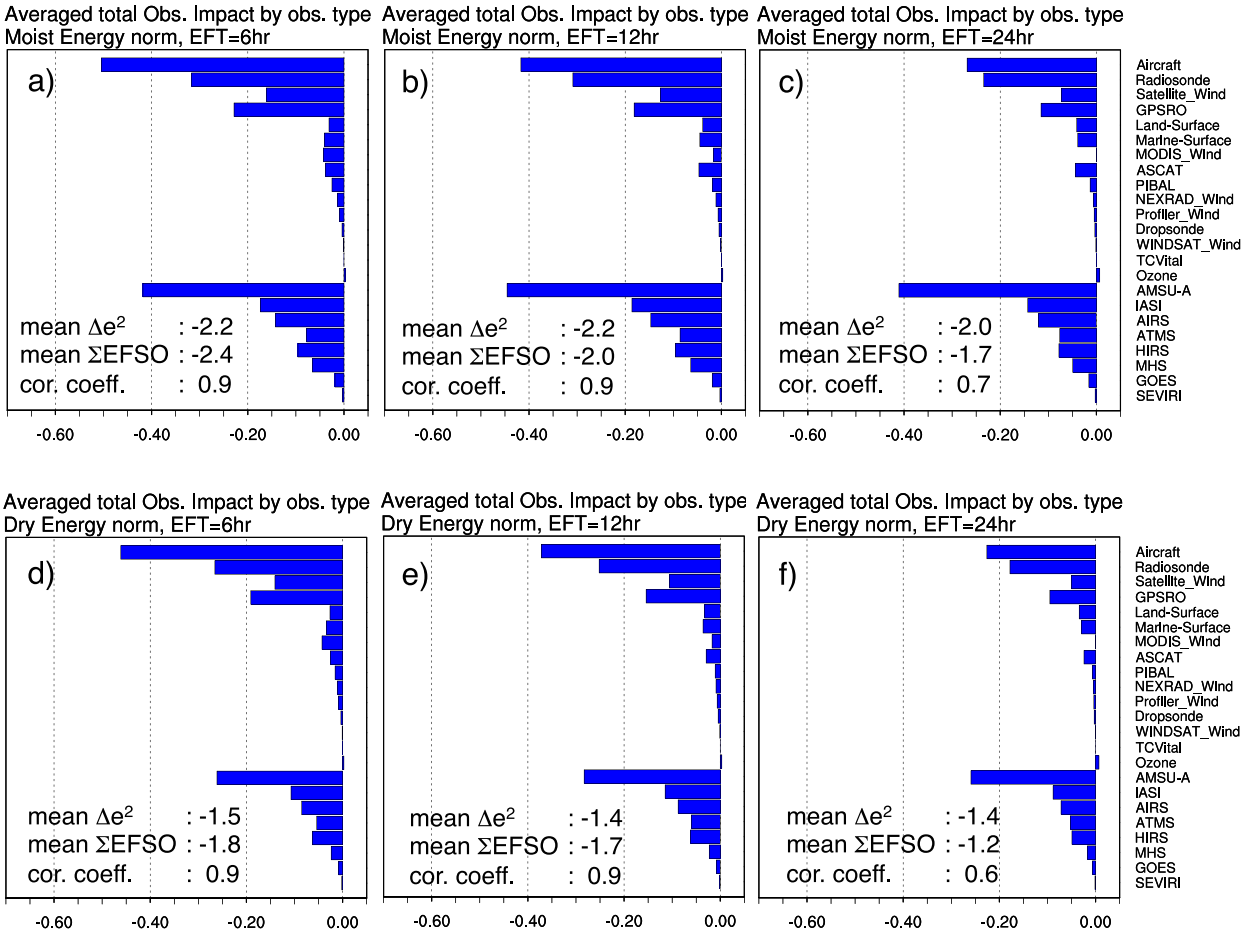

FIG. 1. Comparison of EFSO impacts from each observation type evaluated for different lead times and error norms. (a) The 6-h EFSO impacts measured with the moist total energy norm. (b),(c) As in (a), but for 12-h and 24-h lead times, respectively. (d),(e) As in (a),(b), but measured with the dry total energy norm, respectively. (f) As in (c), but measured with the dry total energy norm. In all panels, the EFSO impacts are verified against the GSI analysis and have units of J kg^{-1}. The mean actual forecast error reduction (mean $\Delta\mathbf{e}^2$), the sum of EFSO impacts over all observations (mean ΣEFSO), and the correlation coefficient of these two quantities (cor. coeff.) over the verification period are also given in each panel.

in particular to what degree it is consistent with the conventionally used 24-h FSO/EFSO.

Each panel in Fig. 1 shows the EFSO impacts from each of the observation types averaged over the 31-day period evaluated with 6- (left panels), 12- (middle panels), and 24-h (right panels) lead times. The top and bottom rows of panels in Fig. 1 show the EFSO impacts measured, respectively, with moist and dry total energy norms. Despite the concern we raised in the introduction, Fig. 1 shows that the EFSO impacts evaluated with different lead times are in fact highly consistent. Two notable features are 1) the estimated impacts decrease as the lead time increases (except for AMSU-A and radiosondes from 6 to 12 h), when, in fact, the forecast error grows, and 2) the decrease in the impacts with lead time is modest for satellite radiances and surface observations (land surface, marine surface, and ASCAT) but is large for other observation types such as aircraft and MODIS winds. Both features could be explained by the inability of the static covariance localization to accurately account for time propagation (Bishop and Hodyss 2009a,b; Gasperoni and Wang 2015): as the lead time increases, the information from an observation is dispersed away from where it was observed, but the localization applied in EFSO fails to accurately account for this effect, resulting in diminished impact estimation. This effect is stronger for longer lead times and in the upper troposphere where the westerly jet prevails. This explains why the impact from aircraft, for example, weakens more quickly than does that from surface observations.

In assessing the accuracy of EFSO impact estimation, it is useful to check the consistency between the actual forecast error reduction [$\Delta\mathbf{e}^2$ as in (4)] and the sum of

the EFSO impacts over all observations. The means and the correlation coefficient of these two quantities over the verification period are also given in each panel in Fig. 1. The correlation coefficients are higher and the means of the two quantities are also more consistent for shorter lead times than for the longer ones, suggesting that EFSO impact estimations are more accurate for shorter lead times. This is consistent with our expectation that EFSO with shorter lead times suffer less from the problem of localization advection.

The EFSO-estimated percentages of beneficial observations in our system for the lead times of 0, 6, 12, and 24 h are shown in Figs. 2a–d, respectively. Consistent with previous studies (see the appendix), the percentage of beneficial observations for 24-h forecasts (Fig. 2d) is only slightly above 50% for all observation types (except TCVital, whose statistics are not reliable as a result of the limited sample size of only 77). Interestingly, however, at shorter lead times, more observations are estimated to be beneficial. The percentages of beneficial observations, all types combined, are 56%, 53%, 52%, and 51%, respectively, for 0, 6, 12, and 24 h. In the appendix, we discuss and interpret these results.

We have seen that, statistically, 6-h EFSO impacts are mostly consistent with 24-h EFSO. We have found that this consistency also holds even for individual observations: for instance, Fig. 3 shows the horizontal (upper panels) and vertical (lower panels) distributions of the EFSO impacts evaluated with lead times of 6 (left panels) and 24 (middle panels) hours for one type of MODIS winds (report type 259) and for one of the identified regional dropout cases (case 17; see Table 2). We are interested in the consistency of the large red circles between the two lead times in terms of their positions because, in PQC, we are concerned only with the $\delta\overline{\mathbf{y}}_0^{\mathrm{ob}}$ innovations with large detrimental impacts. By comparing the left panels with the right panels, we can observe that the observations with large detrimental 24-h EFSO impacts are mostly collocated with those with large detrimental 6-h EFSO impacts, both horizontally and vertically. This visual impression is supported by the scatterplot (not shown) of 6- and 24-h EFSO impacts: the correlation is not very strong near the origin, with some of the innovations with small positive 6-h EFSO having negative 24-h EFSO, but there is a clear positive correlation for the innovations with large EFSO values that we are interested in. Similar results are observed for the other dropout cases.

One could argue that the negatively impacting observations found in Fig. 3 could be simply a result of the failure of the operational QC to reject observations that are inconsistent with other observations or the analysis. As discussed in section 2a, 0-h EFSO can be used to rule out this possibility. By comparing Figs. 3a,b and 3d,e with their 0-h equivalents (Figs. 3c,f), we find that 0-h EFSO are mostly negative, that is, beneficial (blue), which supports the soundness of the operational QC; we can also see from these figures that the large negative EFSO impacts (red circles) found with either 6- or 24-h lead time cannot be detected with 0-h EFSO, which highlights the importance of information from later observations, used in 6- or 24-h EFSO, in detecting the detrimental $\delta\overline{\mathbf{y}}_0^{\mathrm{ob}}$ innovations.

From the scatterplot showing the results between the $\delta\overline{\mathbf{y}}_0^{\mathrm{ob}}$ innovations and the EFSO impacts (not shown), the large negative impacts identified in Figs. 3a,b,d,e are found to be associated with both large innovations $\delta\overline{\mathbf{y}}_0^{\mathrm{ob}}$ and large sensitivities $\partial(\Delta\mathbf{e}^2)/\partial\mathbf{y}$. One may expect that these large impacts should tend to occur when the background field is uncertain and thus show large spread; in fact, we found that observations associated with large background spreads tend to be more detrimental than those with smaller spreads.

c. *Dependence on verifying truth*

As discussed in section 3c, when applying EFSO to a hybrid DA system, the verifying analysis can be taken either from the variational part (GSI analysis) or from the EnKF part (LETKF mean analysis). In general, the variational analysis at higher resolution is considered more accurate; in our particular system, however, LETKF mean analysis has the advantage of being produced at the same resolution as the other variables that appear in (6); being an ensemble average should also be an advantage since small-scale features that are unresolvable with the observing network tend to cancel out. Since it is not clear a priori which of the two analyses is more appropriate, we examined the EFSO's dependence on the choice of verifying truth. The time-averaged EFSO impacts from different types of observation verified against the LETKF mean analysis (not shown) were almost identical to those verified against the GSI analysis, so that the EFSO impacts do not sensitively depend on which verifying truth is used.

EFSO's insensitivity to the choice of verifying truth holds even for individual observations. Scatterplots of the two EFSO values (one verified against GSI analysis and the other against LETKF mean analysis) for different observation types (not shown) clearly show high correlation. With these results, we conclude that the EFSO diagnostics are indeed quite insensitive to the choice of verifying analysis.

The findings in this section lead us to answer affirmatively to questions 1 and 2 posed in the introduction. The fact that both 6- and 24-h EFSOs can consistently identify observations with large negative impacts ensures that 6-h EFSO can be used for PQC. Because the EFSO diagnostics is found not to be too sensitive to the

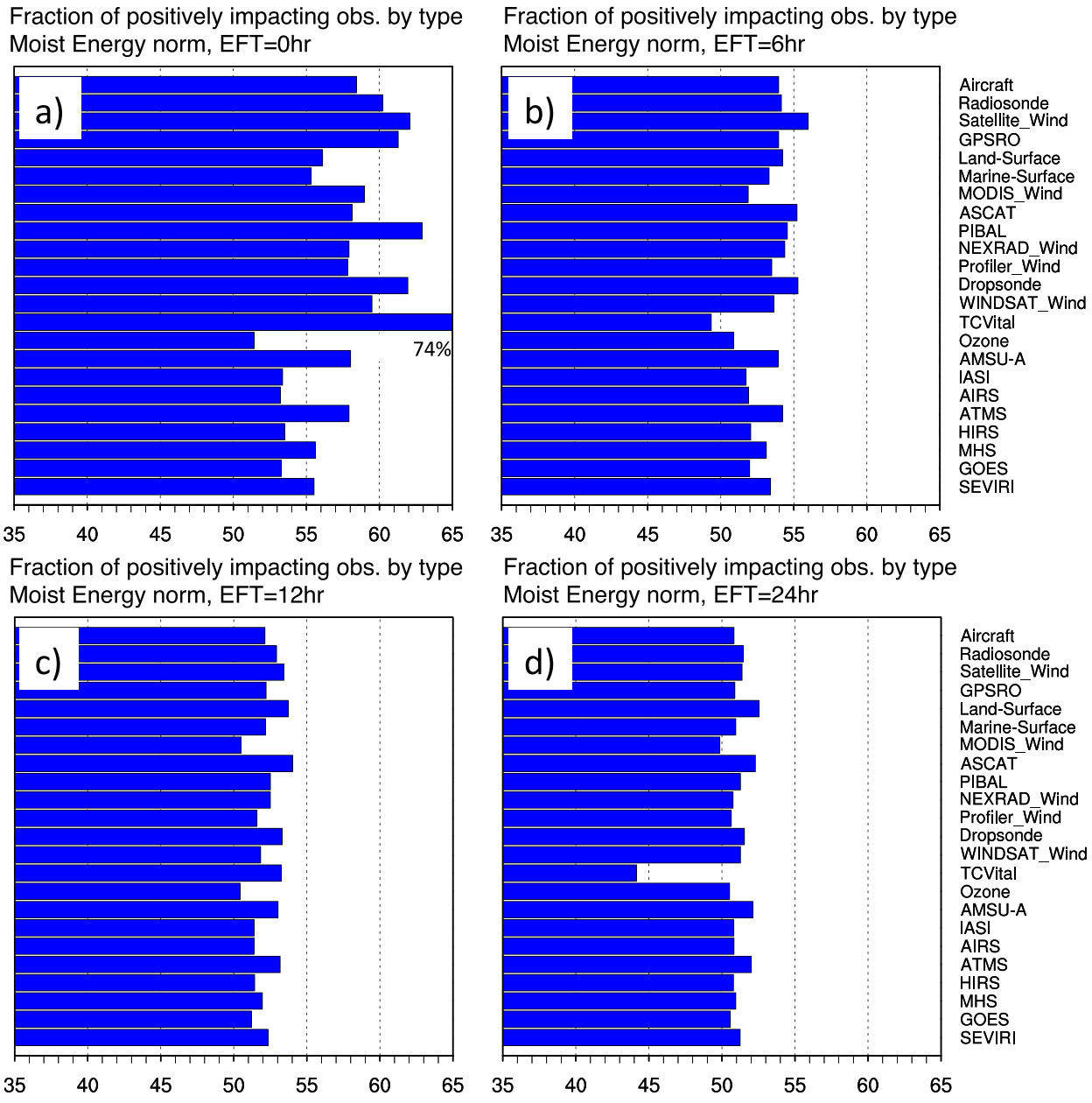

FIG. 2. Percentage of beneficial observations (i.e., the number of observations with positive impacts onto a forecast divided by the number of all observations of the same type and then multiplied by 100) classified by the observation types. EFSO impacts are computed using the moist total energy norm with the control GSI analysis as the verifying truth. Shown are the results evaluated with lead times of (a) 0, (b) 6, (c) 12, and (d) 24 h. Statistics are taken for a 1-month period, with a total observation count of 218 025 941.

choice of verifying truth, in the next section we only show the results verified against the GSI analysis.

5. Detrimental data-denial experiments

In this section we perform a series of detrimental data-denial experiments with different data-denial strategies to answer questions 3 and 4 in the introduction. Here, "detrimental data-denial experiments" denote experiments in which the analyses are repeated without using the detrimental $\delta\overline{\mathbf{y}}_0^{\mathrm{ob}}$ innovations identified by EFSO and the forecasts repeated from the (hopefully improved) new analyses.

a. Selection of cases

Our regional dropout detection criteria (section 2c) resulted in about 200 potential dropout cases. Since our

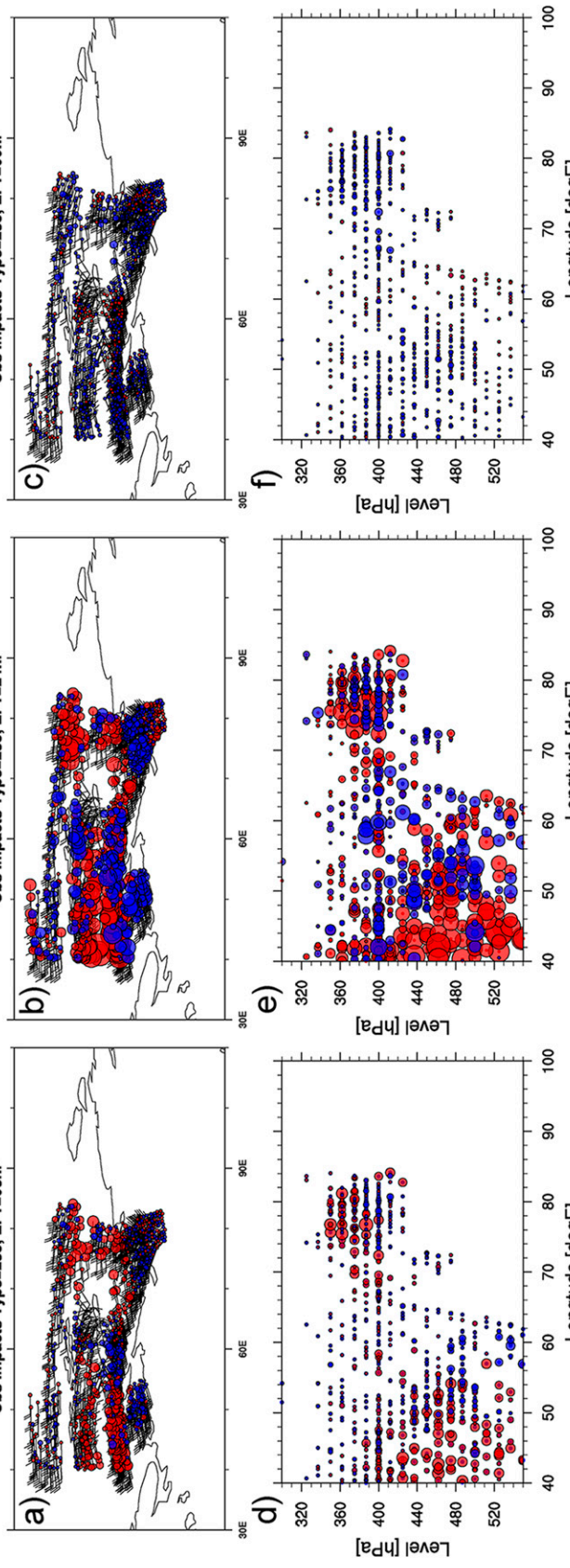

FIG. 3. Geographical and vertical distributions of EFSO impacts for individual MODIS wind (report type 259) observations on one of the regional dropout cases (1800 UTC 6 Feb 2012 initial with the target area of 60°–90°N × 40° –100°E). (a) Horizontal distribution of 6-h EFSO impacts. (b) As in (a), but for 24-h EFSO. (c) As in (a), but for 0-h EFSO. (d) Vertical distribution of 6-h EFSO. (e) As in (d), but for 24-h EFSO. (f) As in (d), but for 0-h EFSO. Red and blue circles represent, respectively, negative and positive impacts (i.e., positive and negative EFSO values). The area of each circle corresponds to the magnitude of the EFSO impact. Wind barbs in (a)–(c) represent the observed wind. Each MODIS wind observation is composed of a pair of observations, one for u (zonal wind) and the other for v (meridional wind), which are assimilated separately. Here, the impact for each MODIS observation is defined as the sum of the impacts from its u and v components. The EFSO is verified against the control GSI analysis and is measured with the moist total energy norm restricted to the abovementioned target area.

TABLE 2. List of the potential regional dropout cases for which data-denial experiments are performed.

				Detection with 6-h EFSO		Detection with 24-h EFSO	
Case No.	Date	Lat	Lon	Detected type	Estimated impact (%)	Detected type	Estimated impact (%)
1	0000 UTC 12 Jan	90°–60°S	110°–140°E	257, 258, and 259 (MODIS)	16	257, 258, and 259 (MODIS)	1.7
2	1800 UTC 12 Jan	60°–90°N	140°E–180°	257, 258, and 259 (MODIS)	20	130 and 133 (aircraft), 257, 258, and 259 (MODIS)	19
3	0600 UTC 13 Jan	60°–90°N	70°–20°W	257, 258, and 259 (MODIS)	30	130, 131, 133, 230, 231, and 233 (aircraft)	34
4	1800 UTC 14 Jan	45°–90°N	120°–150°E	4 (GPSRO), 230 and 231 (aircraft), 257 and 258 (MODIS)	10	4 (GPSRO), 130, 131, 133, 230, 231, and 233 (aircraft)	17
5	1800 UTC 15 Jan	60°–90°N	10°–80°E	120 (radiosonde), 257 and 259 (MODIS)	21	131, 133, 231, and 233 (aircraft)	7
6	1800 UTC 17 Jan	60°–90°N	50°W–0°	4 and 42 (GPSRO), 131 and 133 (aircraft), 257 (MODIS)	18	4, 42, and 722–744 (GPSRO), 130, 131, 133, 230, 231, and 233 (aircraft)	18
7	0600 UTC 18 Jan	90°–60°S	70°–30°W	4 and 42 (GPSRO), AMSU-A and IASI	13	4, 42, and 740–745 (GPSRO), 257, 258, and 259 (MODIS)	25
8	1800 UTC 18 Jan	45°–90°N	120°–150°E	257, 258, and 259 (MODIS)	15	257, 258, and 259 (MODIS)	7
9	1800 UTC 26 Jan	60°–90°N	40°– 80°E	257, 258, and 259 (MODIS)	36	AMSU-A	5
10	0000 UTC 27 Jan	60°–90°N	30°–80°E	4 (GPSRO), 131, 133, and 231 (aircraft), 180 (marine), 257, 258, and 259 (MODIS)	38	131, 133, 231, and 233 (aircraft), 180 (marine), 257, 258, and 259 (MODIS)	22
11	0000 UTC 27 Jan	60°–90°N	20°W–10°E	131 and 230 (aircraft), IASI and AIRS	11	130, 131, 133, 230, 231, and 233 (aircraft)	12
12	1800 UTC 28 Jan	60°–90°N	50°–90°E	180 (marine surface), 257, 258, and 259 (MODIS)	48	257, 258, and 259 (MODIS) HIRS	39
13	1800 UTC 2 Feb	60°–90°N	40°–110°E	120 (radiosonde), 257, 258, and 259 (MODIS)	61	120 and 220 (radiosonde), 257, 258, and 259 (MODIS)	47
14	0000 UTC 3 Feb	60°–90°N	60°–90°E	180 and 280 (marine), 257, 258, and 259 (MODIS)	10	180 (marine), 257, 258, and 259 (MODIS)	11
15	0000 UTC 4 Feb	60°–90°N	40°–10°W	42 (GPSRO), 131 (aircraft)	11	4, 42, 720, 722, and 740 (GPSRO), 120 and 220 (radiosonde), 130, 131, 230, 231, and 233 (aircraft)	48
16	1200 UTC 5 Feb	90°–60°S	60°W–0°	257 and 259 (MODIS)	33	257, 258, and 259 (MODIS)	23
17	1800 UTC 6 Feb	60°–90°N	40°–100°E	120 (radiosonde), 180 (marine surface), 257, 258, and 259 (MODIS)	39	231 (aircraft), 257, 258, and 259 (MODIS)	66
18	1800 UTC 6 Feb	90°–60°S	60°W–10°E	257, 258, and 259 (MODIS), 706–721 (ozone)	23	4 and 740–745 (GPSRO), 257, 258, and 259 (MODIS), 706–721(ozone)	26
19	0600 UTC 9 Feb	60°–90°N	140°–90°W	4 and 42 (GPSRO), 257, 258, and 259 (MODIS), HIRS	47	4, 42, 722, 740, and 744 (GPSRO), AMSU-A, HIRS, IASI, AIRS	42
20	0600 UTC 10 Feb	90°–60°S	50°–80°E	257 and 259 (MODIS)	10	257, 258, and 259 (MODIS)	19

computational resources did not allow us to perform detrimental data-denial experiments on all of them, we picked up top-20 cases that showed the largest *EFSO-estimated forecast improvement* defined as the EFSO-estimated reduction of regional 6-h forecast error normalized by the observed regional 6-h forecast error. The errors are measured using the moist total energy norm. If the target region is formed by merging more than one 30° × 30° cells in the procedure described in section 2c, then the largest value among the unmerged 30° × 30° cells is used for the selection.

The selected 20 cases are summarized in Table 2. The observation types that showed negative net 6-h impacts are shown in the fifth column along with the corresponding "report type" numbers. The EFSO-estimated forecast improvements for each case are shown in the sixth column.

Among the 20 cases listed in Table 2, case 17 deserves special attention because it is exactly the case for which ODKM13 found ~30% regional improvement by rejecting detrimental $\delta\overline{\mathbf{y}}_0^{\mathrm{ob}}$ innovations from MODIS wind observations identified by 24-h EFSO. We document in detail the results for this case in section 5d.

b. Detrimental innovation selection criteria

The *detrimental innovation selection criteria*, which determine which observation to deny given 6-h EFSO for each observation computed with (9) and (10), are a very important component of the PQC algorithm. We begin the exploration of these criteria by critically reviewing the one adopted by ODKM13. Their intricate algorithm implicitly assumes that observations with large negative impact should be clustered in horizontally and vertically localized regions. It is not clear, however, whether such an assumption is justifiable. In fact, as we can see from Fig. 3, the observations with positive and negative impacts are not well separated: for any observation with a large negative impact, we can easily find observations with positive impacts in its vicinity, both vertically and horizontally. Visual inspections for other cases and other observation types (not shown) all support our claim above. We thus conclude that it is more appropriate to choose the observations solely based on their EFSO values rather than to group them based on their geographical and vertical locations.

We now consider how many of the observations to reject given the EFSO values of each observation. To address this issue, we examine the statistical distribution of EFSO values, for each case and for each of the identified detrimental observation types. For each type, we sort the observations in ascending order of their EFSO values and plot the EFSO values against their ranks (the definition of the rank here is such that, if the rank of an observation is r, then there are $r - 1$ observations whose EFSO values are smaller than that of the rank r observation). In choosing the observations to deny, we aim to reduce the forecast errors as much as possible by rejecting as few observations as possible. Thus, if we can find a "jump," that is, a steep slope or a discontinuity at which the EFSO value suddenly becomes large, it seems reasonable to put the threshold there.

Three typical examples of such plots are shown in Fig. 4. In Fig. 4a, we can locate a clear jump near the right edge of the plot. For this kind of distribution, it is easy to choose a threshold. Unfortunately, however, in most of the cases we examined, such clear jumps could not be found. In Figs. 4b,c, the EFSO values are distributed more continuously. For these distributions, it is difficult to objectively determine the best threshold above which the observations can be regarded as "outliers."

Since it is difficult to objectively determine the threshold, we decided to try three simple criteria for determining thresholds and we performed data-denial experiments for each of them. For comparison's sake, we also tried posing no threshold at all; namely, rejecting all observations of the types that are judged by EFSO to be collectively detrimental. We call this criterion allobs. The following list summarizes the three criteria along with the allobs criterion:[2]

- *allobs*—remove all observations of the detected type within the target region, regardless of the EFSO values of each observation;
- *allneg*—remove all negatively impacting observations of the detected type within the target region;
- *one sigma*—remove observations of the detected type within the target region whose EFSO values were above the mean of the same type by at least one standard deviation (σ); and
- *netzero*—for each of the detected observation types, sort the observations within the target region based on the EFSO impacts and remove observations from the one with the largest negative impacts (positive values) until the net impact of that type becomes zero

Of the above four criteria, netzero is the most selective and allobs the least selective. The allobs criterion

[2] To all of the four criteria, we add the following condition: observations located away from the target domain by more than 15° in either latitudinal or longitudinal direction are not rejected. This condition, however, was later found to be unnecessary (redundant) because the localization in (10) automatically makes the impacts of observations away from the target region negligibly small.

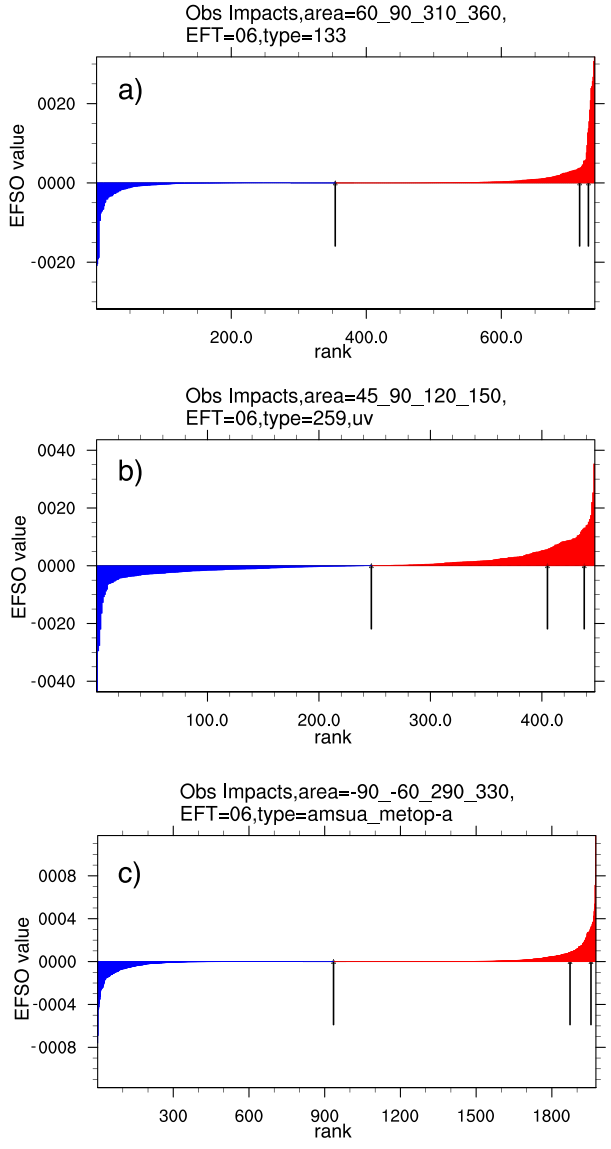

FIG. 4. The 6-h EFSO values (10^{-3} J kg^{-1}) of individual observations plotted against their ranks. Shown are (a) the observations of report type 133 (aircraft) for case 6, (b) the observations of report type 259 (MODIS wind) for case 8, and (c) AMSU-A observations from the *MetOp-A* satellite for case 7. Positive and negative impacts (or negative and positive values, respectively) are plotted with blue and red colors, respectively. The three vertical upward-pointing arrows represent, from left to right, the thresholds for rejection of the allneg, one-sigma, and netzero criteria.

serves as a baseline. The numbers of observations that are denied by each criterion are summarized in Table 3. The allneg criterion denies about half of the observations of detrimental types within the target region, consistent with the fact that about a half of all the observations have negative impacts, but they are still a tiny portion of the total number of the assimilated observations (on the order of 0.1% ~ 1%).

c. *Verification*

With PQC, we aim to locally improve the analysis accuracy to provide an improved first guess to the analysis during the next cycle. It is however very difficult to discern whether an analysis is improved or not. Here, we use the score of a forecast initialized with the analysis in question as a proxy to measure the quality of the analysis, and compare the two forecasts initialized with the analyses, one before and the other after, the denial of detrimental $\delta\overline{\mathbf{y}}_0^{\mathrm{ob}}$ innovations.

We note that this verification approach does not provide compelling proof of analysis improvement since, as has been shown by previous studies, when an analysis is changed to reduce short-term forecast errors, improvement in the forecast at the targeted and longer lead times does not necessarily imply improvement in the analysis in the sense of bringing it closer to the truth. Isaksen et al. (2005) performed a careful examination of whether a perturbation to an analysis based on the adjoint sensitivity technique that enhances a forecast's fit to the observations (or "key analysis errors") can be interpreted as real analysis errors and concluded that such perturbations "cannot justifiably be interpreted as analysis error as far as their detailed structure is concerned," on the ground that 1) the structure of the perturbations depends strongly on the subjective choice of the error norm and 2) the forecast improvements after 12 h and onward are preceded by small but statistically significant forecast degradations at shorter lead times. Kleist and Morgan (2005) pointed out another mechanism where 3) an analysis perturbation that improves a forecast does so by compensating for model errors rather than by reducing the true analysis error. Furthermore, 4) when working with forecast dropouts, it is difficult to rule out the possibility of "regression to the mean," a statistical phenomenon that, in our context, can be stated as "for a forecast dropout (i.e., a case with a poor extreme of forecast skill), any perturbation to the analysis, regardless of its being closer to or farther from the truth, will likely improve the forecast by bringing the forecast skill closer to the population (climatological) mean." While the second issue above seems not to apply to our case, the other three issues may all apply. Hence, the purpose of this verification is modest and the results suggest but do not conclusively prove analysis improvement.

As we are interested in local analysis improvement, we use the metric "local relative forecast improvement" with a relatively short lead time, which is defined as follows.

First, we divide the globe into 10° × 10° patches and compute the t-h forecast error measured with the moist

TABLE 3. The number of rejected observations for each case and each criterion. The percentage with respect to the total number of assimilated observations is shown in parentheses as well.

	Detection with 6-h EFSO			
Case No.	allobs	allneg	one-sigma	netzero
1	1488 (0.07%)	968 (0.05%)	182 (0.01%)	326 (0.02%)
2	2292 (0.13%)	1174 (0.06%)	242 (0.01%)	110 (0.006%)
3	2842 (0.17%)	1714 (0.10%)	224 (0.01%)	344 (0.02%)
4	3827 (0.21%)	2126 (0.12%)	352 (0.02%)	270 (0.02%)
5	3328 (0.19%)	1714 (0.10%)	246 (0.01%)	118 (0.007%)
6	9360 (0.55%)	4430 (0.26%)	230 (0.01%)	22 (0.001%)
7	31491 (1.80%)	15690 (0.90%)	867 (0.05%)	67 (0.004%)
8	3654 (0.20%)	1816 (0.10%)	320 (0.02%)	138 (0.008%)
9	2330 (0.13%)	1510 (0.09%)	296 (0.02%)	510 (0.03%)
10	3278 (0.18%)	1720 (0.09%)	204 (0.01%)	89 (0.005%)
11	27832 (1.50%)	13726 (0.74%)	375 (0.02%)	32 (0.002%)
12	3830 (0.22%)	2282 (0.13%)	526 (0.03%)	462 (0.03%)
13	6416 (0.36%)	3936 (0.22%)	720 (0.04%)	908 (0.05%)
14	481(0.03%)	234 (0.01%)	34 (0.002%)	26 (0.001%)
15	966 (0.05%)	508 (0.03%)	23 (0.001%)	11 (0.0006%)
16	6956 (0.39%)	3544 (0.20%)	522 (0.03%)	174 (0.01%)
17	5915 (0.34%)	3326 (0.19%)	616 (0.04%)	415 (0.02%)
18	6238 (0.36%)	3276 (0.19%)	622 (0.04%)	366 (0.02%)
19	8504 (0.51%)	4678 (0.28%)	749 (0.04%)	809 (0.05%)
20	1216 (0.09%)	598 (0.04%)	128 (0.01%)	48 (0.003%)

norm restricted to that region verified against a GSI analysis that was obtained before applying PQC.[3] The error is computed for each of two forecasts: one initialized by the original analysis and the other by the new analysis obtained by rejecting the identified detrimental $\delta\overline{\mathbf{y}}_0^{\text{ob}}$ innovations, denoted by $\mathbf{e}_{t|0}^{f,\text{beforeQC}}$ and $\mathbf{e}_{t|0}^{f,\text{afterQC}}$, respectively. The relative forecast improvement for each $10^\circ \times 10^\circ$ patch is then defined as

$$\text{relative forecast improvement} := \frac{\mathrm{e}_{t|0}^{f,\text{beforeQC}} - \mathrm{e}_{t|0}^{f,\text{afterQC}}}{\mathrm{e}_{t|0}^{f,\text{beforeQC}}} \times 100(\%). \tag{15}$$

To see PQC's impact at larger scales, we also computed an "average improvement," which is also defined by (15), but for the NH (40°–90°N, 0°–360°) or SH (40°–90°S, 0°–360°) extratropics depending on where the target region is.

d. Case study 1: A typical successful case (case 17)

We now proceed to show the results of data-denial experiments. First, we show the results for case 17, for which ODKM13 obtained particularly successful results. The 6-h EFSO impacts estimated for each observation type for case 17 are shown in Fig. 5a. Consistent with ODKM13, MODIS winds were clearly identified as detrimental. In addition, radiosonde and marine-surface observations were also found to have a slightly negative net impact.

The relative forecast improvements for case 17 by the data denial based on 6-h EFSO are summarized in Fig. 6. The allobs column shows that if we deny all observations of the types judged detrimental by 6-h EFSO, the forecast is improved (blue shades) in some regions but is also degraded (red shades) in other regions, although the degraded regions become smaller as the lead time increases. Thus, rejecting all observations of detrimental type regardless of the EFSO values of the individual observations is not a good strategy. If we remove only the observations that had negative impacts (allneg), we can effectively eliminate most degradation, and the forecast improvement

[3] In our experimental setup, two different analyses are available that are obtained before and after applying PQC. The PQC procedure reduces the number of assimilated observations, resulting, in general, in smaller analysis increments. Thus, if the forecast is verified against the analysis after PQC, the forecast from the analysis after PQC would automatically yield smaller forecast errors than that without PQC. Here, to avoid such a biased judgement, we use the analysis before PQC as the verifying truth.

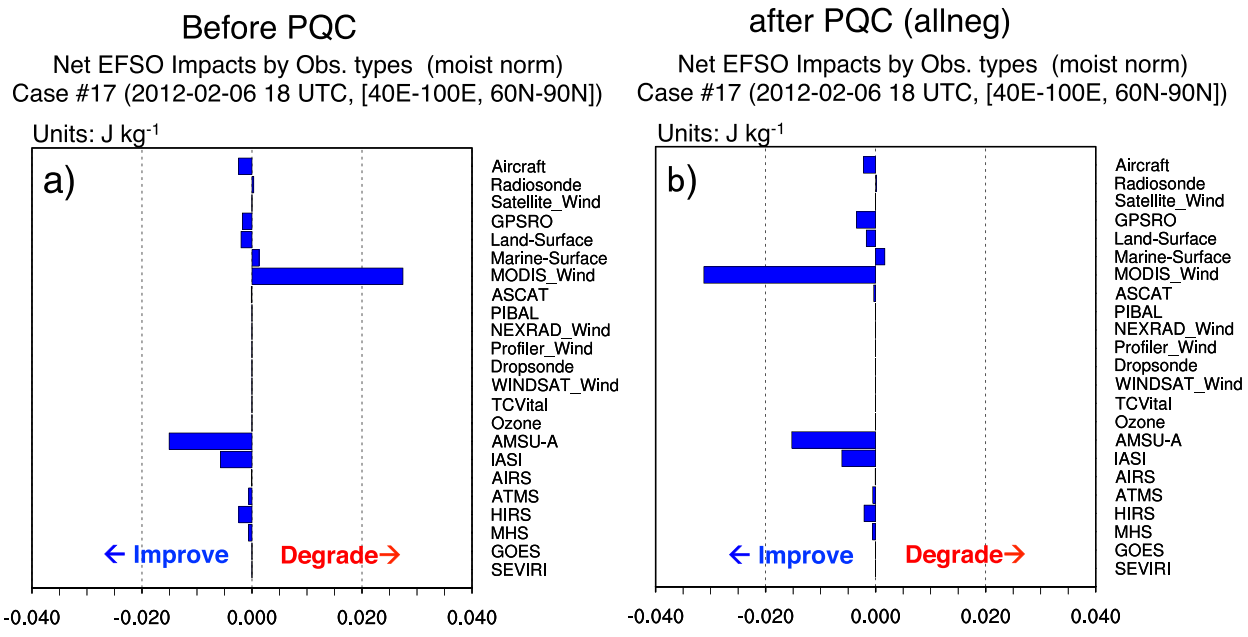

FIG. 5. The regional 6-h forecast error reduction or increase attributed by EFSO to each observation types for case 17, evaluated for the experiments (a) before applying PQC and (b) after applying PQC with the allneg criteria. The forecast errors are measured with the moist total energy norm (J kg^{-1}) within the target area of 60°–90°N, 40°–100°E.

locally reaches as much as 48% for a 24-h forecast. The fact that allneg yields better improvement than allobs while dramatically reducing degradation proves the effectiveness of the EFSO diagnostics. By further restricting the denied observations (one sigma), we can further eliminate forecast degradation; however, this is achieved at the expense of diminished forecast improvement. If we restrict the denied observations even more (netzero), the forecast improvement becomes even smaller.

To check whether the denial of detrimental $\delta\overline{\mathbf{y}}_0^{\mathrm{ob}}$ innovations improved the observational impact on a forecast as expected, it is instructive to inspect the EFSO impacts of the remaining observations using the dataset obtained after PQC. Figure 5b shows that the data denial with the allneg criteria did indeed improved the observational impacts as expected, rendering MODIS winds, which were detrimental before PQC, the most beneficial type. Similar improvements of impact from the denied types of observations were also seen in other dropout cases examined.

e. Case study 2: The most unsuccessful case (case 5)

In the next section, we show that in 18 out of 20 cases, we can in fact improve the forecast by PQC based on 6-h EFSO. However, in two cases, 5 and 9, the denial of observations based on 6- or 24-h EFSO failed to improve the forecast, so we examine in Fig. 7 the results for case 5 (case 9, not shown, exhibited similar features). Looking at the first row (FT = 06), we observe that, with the allneg, one-sigma, and netzero criteria, the 6-h forecast is actually improved within the target area. Thus, the 6-h EFSO is actually accurate in the sense that the EFSO impact and the actual nonlinear impact are consistent. However, beyond 6 h, the forecast improvement almost disappears. We speculate that, in these unsuccessful cases, the detrimental impacts from the denied observations project strongly on decaying modes, so that the effect of not using them eventually diminishes (see the discussion in the appendix). Why we have a few unsuccessful cases needs to be explored in our future work.

f. Summary of the 24-h forecast results

In Table 4 we show, for each case and data-denial strategy, the largest local relative improvement (“max”) and degradation (“min”) of a 24-h forecast and the average hemispheric-scale 24-h forecast improvement evaluated for the extratropics of the hemisphere in which the target region is located (“avg”). For the allneg, one-sigma and netzero criteria based on 6-h EFSO, the average improvement (avg) is positive in almost all cases. The only exceptions are for allneg in cases 5 and 9, where we had degradations, respectively, of 0.2% and 0.4%. Because PQC is designed to minimize the occurrences of local forecast failures, its impact is spatially localized and thus the impact becomes small if averaged over a large spatial domain. This is why the hemispheric-scale average improvement is at most ~2%. We point

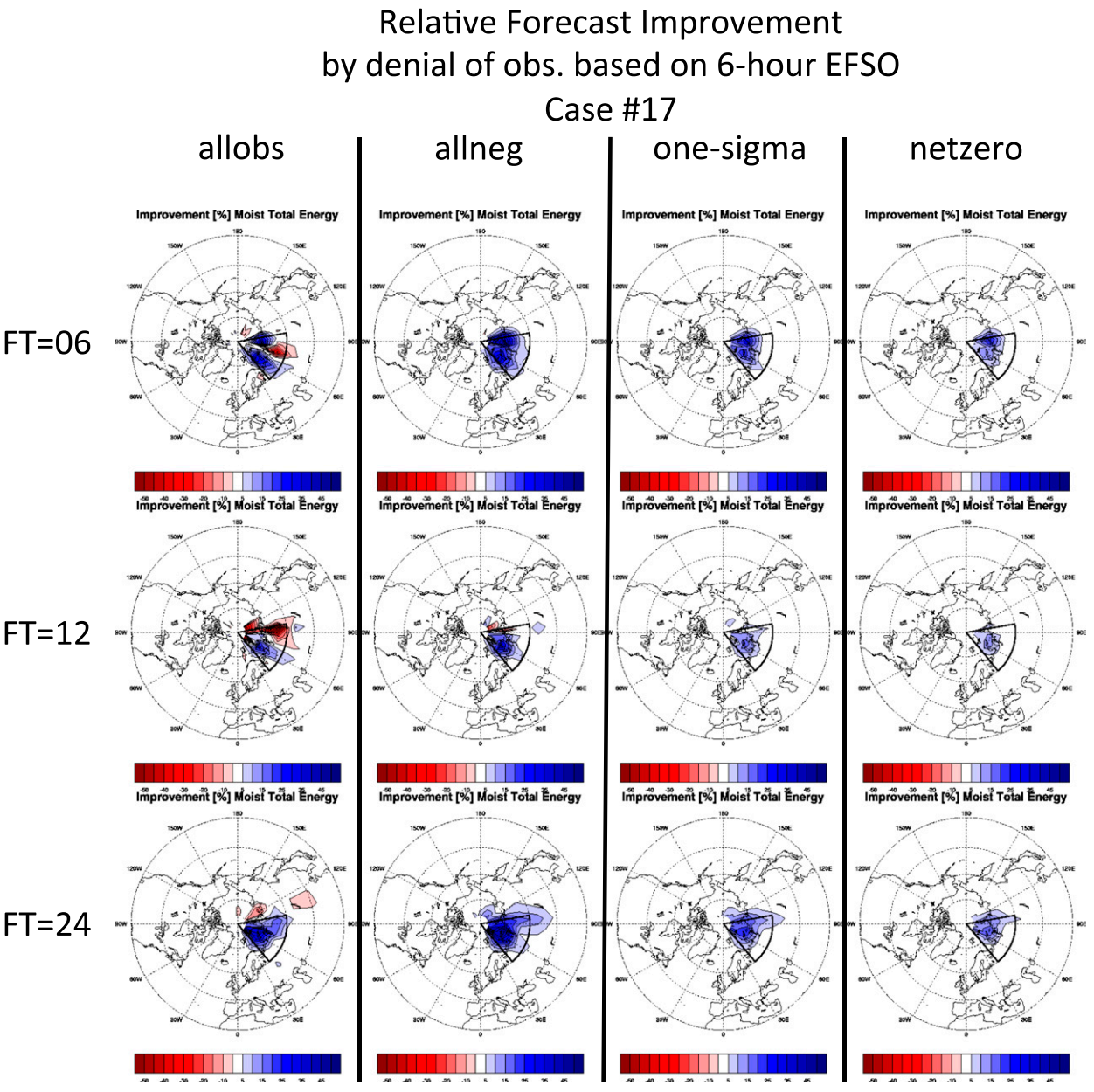

FIG. 6. Relative forecast improvements for each of the four data rejection criteria based on 6-h EFSO for case 17. Each column represents, from left to right, the allobs, allneg, one-sigma, and netzero criteria. The first row represents the relative improvement of 6-h forecasts; the second and third rows represent, respectively, the improvement of 12- and 24-h forecasts. Red (blue) colors represent forecast degradation (improvement). The thick black triangular sector represents the target region. The color bars run from −50 on the left to −10 and on the right from +5 to +45 in increments of 10.

out nevertheless that, although the improvement on the order of ~0.2% to ~2% might seem to be modest, it is normally very difficult to obtain.

The features that we saw for case 17 are also valid for most other cases, namely 1) allobs exhibits both improvement and degradation, 2) allneg alleviates the degradation seen in allobs and tends to show larger improvement, and 3) one-sigma and netzero further reduce the degradation but with reduced improvement.

Encouragingly, the large forecast improvements that we saw for case 17 are not limited only to this particular case. For example, if we look at the allneg criterion, the cases 8, 12, 13, 16, 17, 18, and 19 all exhibit local maximum forecast improvement that exceed 30%. For these cases, the one-sigma and netzero criteria also result in

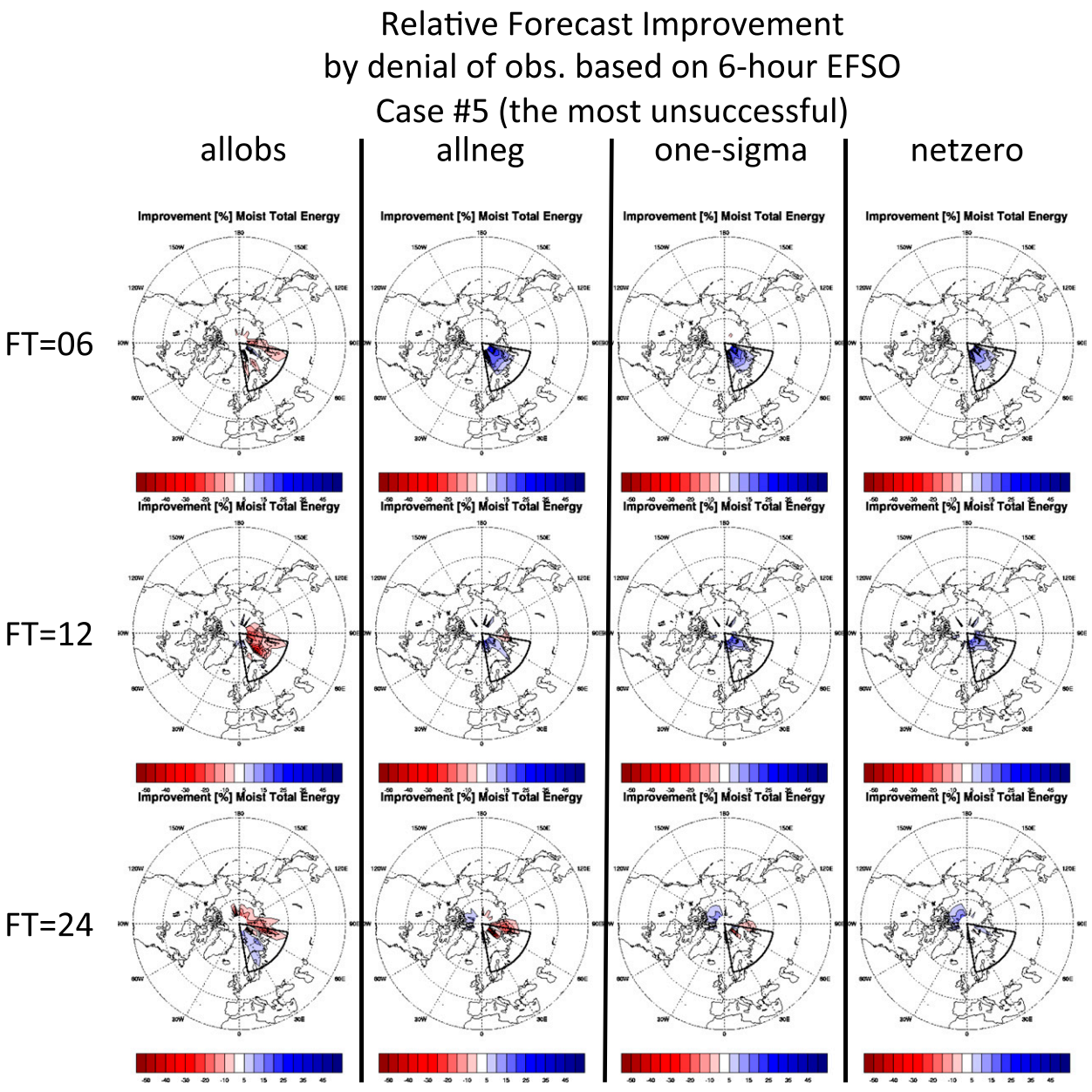

FIG. 7. As in Fig. 6, but for the most unsuccessful case (case 5).

large maximum local forecast improvements (~20%). For all of these particularly successful cases, EFSO identified MODIS wind as the detrimental observation type. This suggests that either the observations from MODIS winds had anomalous errors or the way the DA system handled them was faulty.

g. *Impacts on 5-day forecasts*

In the previous sections, we showed that the strategic denial of observations under the guidance of 6-h EFSO improves the 24-h forecast, which suggests that PQC improved the analyses. To further investigate the benefit of using PQC, we now explore the temporal extension of the forecast impact of PQC.

We explored the ability of PQC to reduce the relative integrated forecast error within three latitudinal bands: the NH extratropics, the SH extratropics, and the tropics (30°S–30°N, within which we made no detrimental innovation denial). These statistics were computed for all 20 cases using the allneg criterion (section 5b). The PQC-modified forecasts from 6 to 126 h were verified against the GSI analysis that was obtained before

TABLE 4. Relative improvement or degradation of 24-h forecasts (%) by the denial of observations based on 6-h EFSO.

	allobs			allneg			one-sigma			netzero		
Case No.	Max	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max	Min	Avg
1	12	−9	0.0	11	−1	0.2	4	−1	0.1	5	−1	0.1
2	14	−5	−0.1	11	−4	0.3	8	−2	0.2	4	0	0.2
3	13	−15	0.0	7	−5	0.2	2	−1	0.0	4	−2	0.0
4	25	−5	0.6	27	−5	0.7	15	−2	0.3	13	−2	0.2
5	15	−32	−0.2	19	−81	−0.2	23	−30	0.2	22	−13	0.3
6	9	−9	0.0	15	−6	0.4	12	−3	0.3	3	−1	0.1
7	17	−9	−0.0	13	−5	0.4	2	−3	0.0	0	0	0.0
8	41	−18	0.9	41	−14	1.1	21	−5	0.8	10	−2	0.4
9	7	−21	−0.6	8	−16	−0.4	8	−3	0.0	8	−4	0.1
10	25	−6	1.1	19	−6	0.7	3	−2	0.2	6	0	0.2
11	11	−6	0.5	9	−5	0.3	2	−2	0.1	3	0	0.1
12	37	−14	0.7	39	−12	0.7	19	−2	0.5	19	−2	0.5
13	24	−9	1.4	30	−9	0.8	18	−10	0.3	19	−12	0.4
14	5	0	0.3	3	0	0.1	1	0	0.0	1	0	0.1
15	3	−2	0.1	1	−1	0.1	1	−1	−0.0	1	−1	0.0
16	27	−15	1.9	30	−21	1.8	23	−4	1.3	16	−2	0.7
17	39	−15	0.8	48	−4	2.1	26	−2	1.2	20	−2	0.8
18	46	−9	2.4	46	−8	2.2	25	−3	1.0	21	−2	0.8
19	44	−24	2.2	37	−10	2.2	17	−1	1.0	14	−1	1.0
20	12	−3	0.2	10	−1	0.3	5	−1	0.2	3	−1	0.0

applying PQC. The forecast errors were measured with the moist total energy norm. In exploring the impact of PQC on 5-day forecasts, when cases of detrimental $\delta\overline{\mathbf{y}}_0^{\text{ob}}$ innovations took place at the same time (such as cases 17 and 18), they were combined at the observation denial step (the fifth step in the PQC algorithm), as they would be in operations. In practice, it is necessary to avoid performing PQC on insignificant cases in order to minimize computational costs and prevent the introduction of noise that can potentially degrade forecasts. Hence, a 20% threshold of 6-h EFSO estimated improvement in the target area (the sixth column in Table 2) was set to distinguish the 11 ''significant'' cases from the other 9 ''insignificant'' cases.

We show the average 5-day forecast error improvement of both the 11 significant cases (Fig. 8a) and the other 9 insignificant cases (Fig. 8b). Since all the regional dropouts and the associated denied observations were in higher latitudes, the improvement takes place mostly in the NH and SH extratropics. The forecast error in the tropical belt, on the other hand, is apparently worse after PQC in the first 6 h, but this degradation is due to the analysis change introduced by PQC rather than to a true degradation, as shown by the fact that it vanishes within 12 h. In contrast, at higher latitudes, where the detrimental $\delta\overline{\mathbf{y}}_0^{\text{ob}}$ innovations were actually denied, the initial improvement persists and grows with time. The growth of improvement can also be seen in the global mean. It reaches about 1% after 5 days, demonstrating the long-term effect of PQC. It is noteworthy that the improvements in higher latitudes tend to expand in space and ''leak'' to lower latitudes (not shown), leading to a steady growth of improvement in the tropics for the 11 significant cases.

The average of the nine insignificant cases in Fig. 8b shows a different scenario. Although the forecasts in the extratropics were still improved until day 4, this improvement was smaller and reached only 0.5%. In addition, the PQC-modified forecasts started degrading after 4 days in most regions.

h. Comparison with data selection based on 24-h EFSO

In the preceding sections, we have confirmed strong evidence of the capacity of PQC based on 6-h EFSO to improve our analysis in regional dropout cases. One could argue, however, that data selection based on 24-h EFSO might result in even better improvement. To see if this is true, we have repeated all the data-denial experiments using 24- instead of 6-h EFSO in the detrimental innovation selection criteria. The observation types that showed negative net 24-h impacts and the *EFSO-estimated forecast improvements* evaluated with 24-h lead time are shown, respectively, in the seventh and eighth columns of Table 2. We found good agreement between the types of observations identified as detrimental by 6- and 24-h EFSOs. The forecast improvements achieved by PQC based on 6- and 24-h EFSOs were also very close. Detailed inspection of the patterns of forecast improvements similar to the plots in

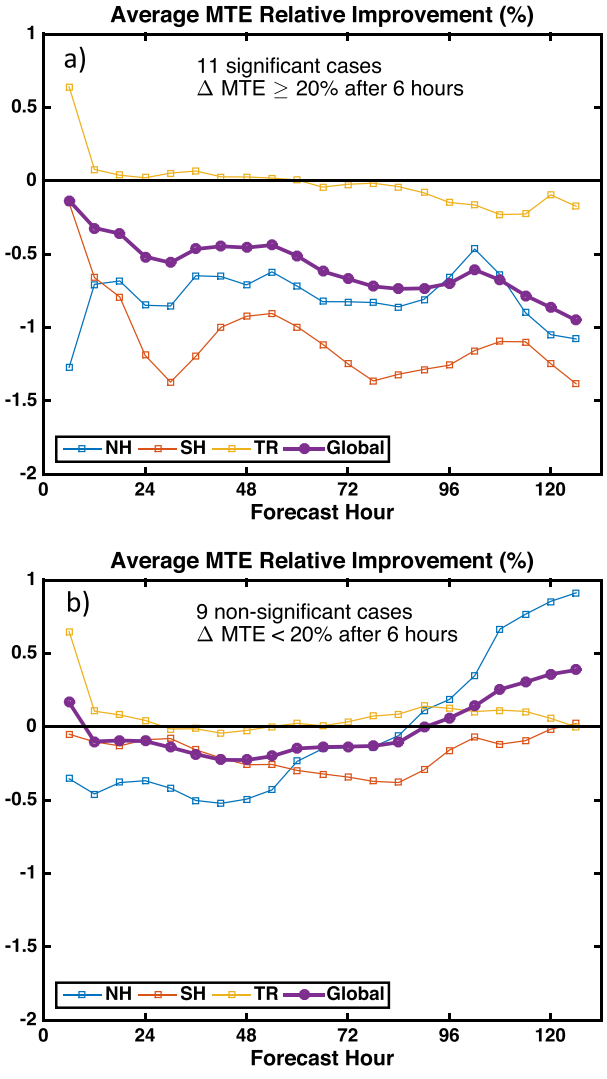

FIG. 8. Relative forecast improvements measured with the moist total energy norm averaged over (a) the 11 significant cases for which the EFSO-estimated 6-h forecast improvement for the target domain exceeded 20% and (b) the 9 nonsignificant cases. The relative forecast improvements are computed for each of the NH extratropics (30°–90°N), the SH extratropics (90°–30°S), the tropics (30°S–30°N), and the whole globe: NH (blue), SH (brown), TR (yellow), and Global (purple).

Fig. 6 (not shown) confirms that the areas of forecast improvement or degradation are similar, regardless of the choice of lead time in EFSO estimation. This corroborates our finding that EFSO is insensitive to the choice of lead time (section 4b).

The results of data-denial experiments with PQC based on 24-h EFSO confirm that PQC with 6-h EFSO is similarly able to improve the forecast and thus the analysis as PQC with 24-h EFSO, but the use of a shorter lead time is much more advantageous for operational implementation.

6. Discussion and conclusions

Recent studies have shown that many "forecast skill dropouts" are due to deficiencies in the initial conditions, not in the model, suggesting that one possible way to mitigate the dropout problem is to improve the operational QC system. In this paper we proposed a new, fully flow-dependent QC technique based on EFSO, which we call proactive QC (PQC), and investigated its effectiveness and feasibility within a quasi-operational context.

A critical assumption behind the concept of PQC is that 6-h EFSO is capable of detecting detrimental $\delta\bar{\mathbf{y}}_0^{\mathrm{ob}}$ innovations. Using a lower-resolution version of NCEP's operational global NWP system, we have confirmed that EFSO results are indeed rather insensitive to the choice of forecast lead time and verifying truth. We then investigated the effectiveness of PQC based on 6-h EFSO by performing data-denial experiments with 20 notable *regional dropout* cases. We found that, by rejecting all the negatively impacting observations (in terms of 6-h EFSO) of the types identified as collectively detrimental (the allneg criterion), the 24-h forecasts were improved in 18 out of 20 cases, with local forecast improvements reaching over 30% in as many as seven cases. Even more encouragingly, the positive impact of PQC on forecasts was found to persist beyond 5 days. As discussed in section 5c, the verification approach adopted here has several limitations and the results do not conclusively prove analysis improvement, but are nevertheless suggestive of our assertion that the analyses can be improved by PQC.

We examined the forecast improvement by PQC to show that the analysis was likely improved so that the 6-h forecast from this analysis, or the first guess at the next analysis cycle, also likely improved. The forecast improvement presented in section 5 does not guarantee forecast improvements within a real-time operational context, because PQC requires the analysis at the later cycle that, on average, produces better forecasts than the one initialized with the PQC-improved previous analysis [i.e., $(t-6)$-h forecast $\mathbf{x}_{t|6}^{f}$ is more accurate than t-h forecast $\mathbf{x}_{t|0}^{f,\mathrm{afterQC}}$]; it is by providing a better first guess to this next analysis and doing it anew that PQC may improve real-time operational forecasts, as discussed later in this section.

An application for which PQC can be used most straightforwardly is reanalysis because, for this non-real-time application, having to wait 6 h is not an issue. We assert, however, that PQC can still be used in real-time operational systems provided that several technical issues are addressed, as discussed below.

The first and perhaps most critical concern is the cost. It would seem that having to wait until the completion of

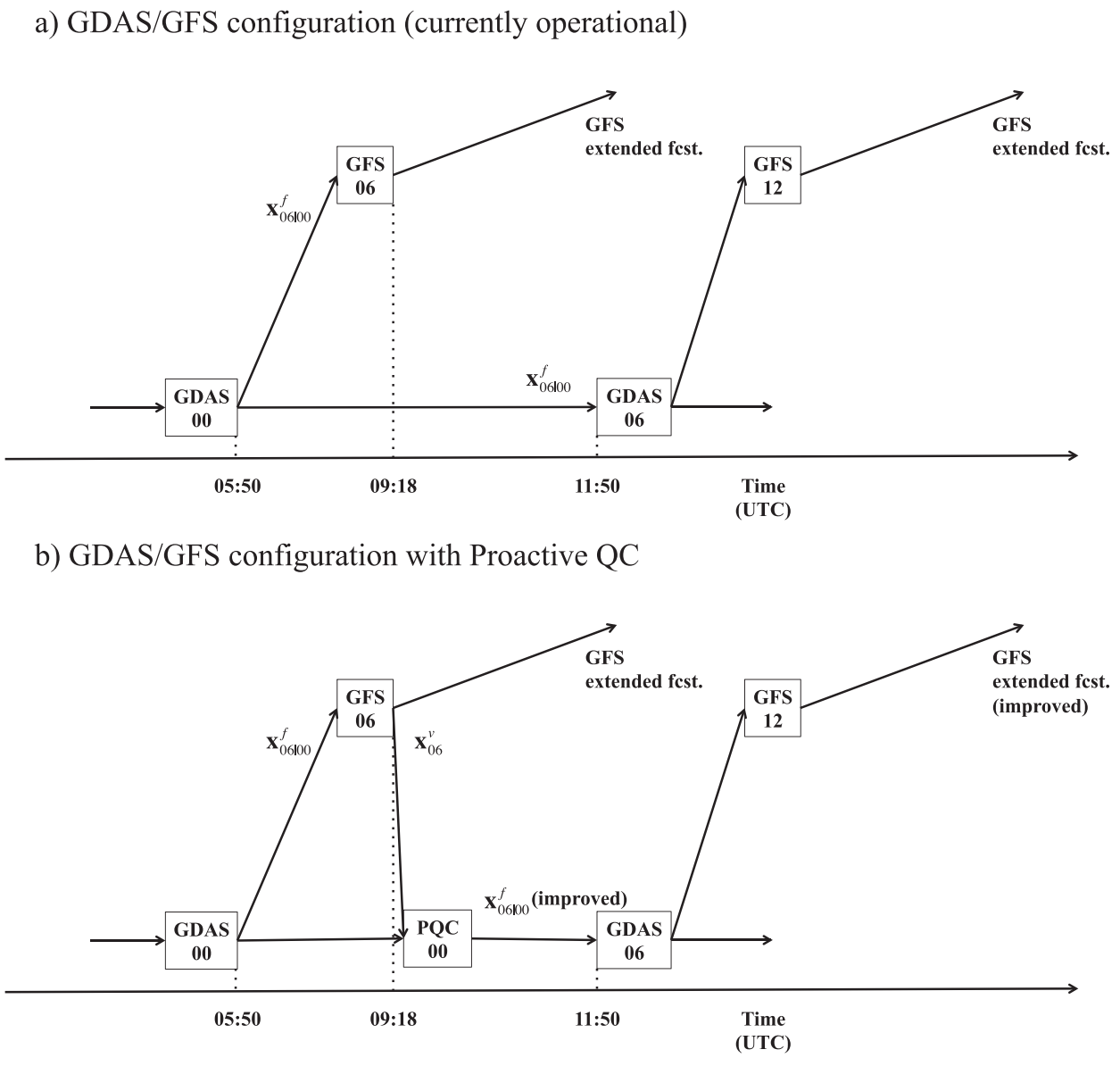

FIG. 9. Schematic illustrations of the early analysis (GFS) and final analysis (GDAS) configurations for 0000–0600 UTC initializations: (a) the current operational configuration and (b) the configuration with proactive QC.

the next analysis would be a serious bottleneck in real-time operations; furthermore, having to perform again the analysis might seem prohibitive. In fact, however, PQC can be performed virtually with no additional cost to the operational system; we now present two ideas that allow for the economical execution of PQC:

One idea (J. Derber, NCEP, 2013, personal communication) is to exploit the time lag between the "early" and "final" analyses: most operational NWP centers, including NCEP, ECMWF, and JMA, maintain two different kinds of DA jobs. The early analysis (called GFS analysis at NCEP) adopts a shorter cutoff time to provide initial conditions for the forecasts in a timely manner, allowing fewer observations to be assimilated. Thus, the early analysis is not carried over to the next cycle. On the other hand, the final analysis (called GDAS analysis at NCEP) adopts a longer cutoff time to allow for the late arrival of observations, and the resulting analysis is taken over to the subsequent cycles. A typical configuration of an operational DA system is illustrated schematically in Fig. 9a, taking NCEP's GDAS/GFS as an example. This type of configuration, common to many operational NWP systems, allows PQC to be executed without any delay to the schedule

(Fig. 9b): suppose, for example, we wish to apply PQC to the final analysis for 0000 UTC (GDAS00) to provide a better background to the final analysis at 0600 UTC (GDAS06). Thanks to the early analysis for 0600 UTC (GFS06), which finishes at 0918 UTC in real time, the verification state $\mathbf{x}^v_{06}$ required to run PQC for the 0000 UTC observations becomes available 2.5 h before GDAS06 starts at 1150 UTC in real time. This time slot of 2.5 h can accommodate well the execution of PQC, providing an improved background $\mathbf{x}^f_{06|00}$ to GDAS06 in a timely manner. The improved background $\mathbf{x}^f_{06|00}$ will result in improvement of the GDAS06 analysis and the subsequent extended forecast at GFS12. For this PQC application, all that is required is that the background $\mathbf{x}^f_{06|00}$ be improved.

One may still argue that the time available for PQC may not be long enough for running the analysis and forecast again. Our second idea addresses this concern. Once we have the list of observations that should be rejected, we can approximately obtain the improved analysis with (12) and (13) in section 2a, which require only a minimal computational cost. This technique enables a cost-efficient estimation of how the analysis and forecast would change by not assimilating the detrimental $\delta\overline{\mathbf{y}}^{\mathrm{ob}}_0$ innovations without actually repeating the analysis and forecast. This approximation in (12) and (13) should be valid if the number of rejected observations is much smaller than that of all the assimilated observations, a condition that was satisfied in all 20 of the cases examined in this paper (cf. Table 3) and should be satisfied in virtually any PQC applications, if the impact of the use of inflation in the EnKF is accounted for. In fact, ODKM13 used the same approximation and obtained very good consistency between the actual nonlinear forecast change and its linear "constant $\mathbf{K}$" approximation (see their Fig. 9).

The second concern is whether the forecast is also improved in an "on line" cycled environment; the data-denial experiments we conducted are "off line" in the sense that the improvement of the forecast achieved by PQC was not taken over to the next cycle. However, if PQC is implemented in real-time operations, the improved forecast will be used as the background at the next cycle. This cycling should not degrade the forecast or reduce the forecast improvement; on the contrary, the improvement should accumulate over the cycles. However, this accumulation of improvements has to be established before bringing PQC into operations.

In the previous section, we argued that the fact that the rejection of some of the MODIS wind observations resulted in particularly large forecast improvements suggests that MODIS wind observations or the corresponding background might have had some technical issues, either in the dataset itself or in the way the data are processed by the NWP system. This aspect deserves a careful examination. While it is logically incorrect to suspect quality issues in either observations or background by the mere evidence of large negative EFSO impacts (section 1, paragraph 5), it should be legitimate to assume that flaws in the observations, the background, or the DA system, if present, will likely manifest themselves as detrimental EFSO impacts, although they will be subject to stochastic fluctuations. Detrimental EFSO impacts would therefore provide useful guidance on which cases to examine to identify potential flaws in the NWP system or observations. This motivates us to explore another major application of EFSO, in addition to improving the analyses and forecasts by PQC. Real-time operation of PQC would enable building a detailed database of EFSO impacts along with relevant metadata as its by-product. Such database of EFSO and metadata can then be provided to instrument/algorithm developers, NWP data specialists, and modelers to help them to identify the problem that produced the negative impacts and avoid them in the future. For this application, close collaboration with instrument/algorithm developers, modelers, and data assimilation specialists is indispensable in determining what type of information and metadata would be most helpful to them. Such collaboration is taking place with the MODIS wind algorithm developers at Cooperative Institute for Meteorological Satellite Studies (CIMSS) at the University of Wisconsin–Madison, and is providing useful information about the biases in MODIS wind O-B innovations.

Finally, we point out that (E)FSO would also allow a more efficient and precise determination of the optimal way to assimilate new observing systems. The current standard observing system experiment (OSE) approach has difficulties in obtaining statistical significance in the presence of the already assimilated observations (Geer 2016). An (E)FSO approach should address this problem by finding the short-term impact of each observation and allowing comparison of the impact of different preprocessing algorithms in a more statistically consistent manner.[4] Lien (2014) already showed that EFSO can be effectively used to systematically design a data-selection strategy, using the TRMM-retrieved global precipitation as an example of a new observing

[4] Lorenc and Marriott (2014) found, when applying FSO to a toy model, that 183 independent cases were needed to discern the impact from a *single observation* with 95% statistical significance; in contrast, Geer (2016) reports that, with an OSE approach, as much as 424 independent forecasts are necessary to discern the *collective* impact of *all observations from an AMSU-A instrument* at the 95% significance level.

system, and avoiding having to perform a large number of expensive OSEs in order to arrive to an appropriate data-selection/QC strategy.

Acknowledgments. This work is based on DH's Ph.D. dissertation, which was supported by the Japan Meteorological Agency (JMA) through the Japanese Government Long-term Overseas Fellowship Program. The computational resources were generously provided by the Joint Center for Satellite Data Assimilation (JCSDA) through their Supercomputer for Satellite Simulations and Data Assimilation Studies [the "S4 supercomputer"; Boukabara et al. (2016)]. We express our gratitude to Dr. Sid Boukabara for this support and to Dr. Jim Jung for his essential guidance on the use of the S4 supercomputer. Constructive and critical comments from Prof. Andrew Lorenc and two anonymous reviewers significantly enhanced the manuscript, for which the authors are most grateful. This work was partially supported by a NOAA/NESDIS/JPSS Proving Ground (PG) and a Risk Reduction (RR) CICS Grant NA14NES4320003.

APPENDIX

An Interpretation of Fig. 2

In the literature on FSO studies, there have been several discussions about what percentage of the observations has a beneficial impact on the forecasts, in particular why so few do, because past FSO studies report that only slightly more than 50% of observations have positive FSO impacts. Recently, Lorenc and Marriott (2014) presented a review of the different views offered on this issue by previous studies and showed, through a series of idealized experiments discussing not only the suboptimality of the DA system (Gelaro et al. 2010) and the limited accuracy of verifying truth (Daescu 2009) but also the differences in the growth rates of each mode of the forecast model, along with the lack of flow dependence in the prescribed **B** matrix, all of which contribute to the lowered fraction of beneficial observations. In view of these discussions, it is interesting to see how the percentage of beneficial observations changes with the evaluation lead time in our system. As we discussed in section 4b, we found that only slightly more than 50% of the observations contribute positively to the forecast at 24-h lead time, but this ratio becomes larger as the lead time gets shorter. Following the arguments of Pires et al. (1996), Trevisan and Uboldi (2004), Uboldi and Trevisan (2006), Carrassi et al. (2007), Trevisan et al. (2010), and Lorenc and Marriott (2014), we can interpret this as follows.

The atmosphere as a dynamical system has both growing and decaying modes. Assume that an observation improves the analysis by significantly improving the decaying modes, but, at the same time, it slightly degrades the growing modes. For a very short forecast, this observation would show a beneficial impact. With time, however, the small initial amplitude in the growing modes will amplify and overwhelm the reduction of the error in the decaying modes, rendering the net impact of that observation negative.

The assimilation in unstable subspace (AUS) approach (Trevisan and Uboldi 2004; Uboldi and Trevisan 2006; Trevisan et al. 2010) restricts analysis increments in the unstable subspace of the model dynamics. Cycled over a long enough period, and under a perfect-model assumption, this method makes the true background errors project entirely on growing (and neutral) modes. In such a system, more observations will exhibit beneficial forecast impacts than in a conventional DA system (Lorenc and Marriott 2014), allowing for a more effective use of the observations. It remains to be investigated, however, whether this advantage of the AUS approach, at present only verified for an idealized toy system with an identical-twin setup, holds also for realistic NWP systems where biases constitute a substantial fraction of the background and observation errors.

REFERENCES

Alpert, J. C., D. L. Carlis, B. A. Ballish, and V. K. Kumar, 2009: Using pseudo RAOB observations to study GFS skill score dropouts. *23rd Conf. on Weather Analysis and Forecasting/19th Conf. on Numerical Weather Prediction*, Omaha, NE, Amer. Meteor. Soc., 5A.6. [Available online at https://ams.confex.com/ams/23WAF19NWP/techprogram/paper_154268.htm.]

Anderson, E., and H. Järvinen, 1999: Variational quality control. *Quart. J. Roy. Meteor. Soc.*, **125**, 697–722, doi:10.1002/qj.49712555416.

Bishop, C. H., and D. Hodyss, 2009a: Ensemble covariances adaptively localized with ECO-RAP. Part 1: Tests on simple error models. *Tellus*, **61A**, 84–96, doi:10.1111/j.1600-0870.2008.00371.x.

——, and ——, 2009b: Ensemble covariances adaptively localized with ECO-RAP. Part 2: A strategy for the atmosphere. *Tellus*, **61A**, 97–111, doi:10.1111/j.1600-0870.2008.00372.x.

Boukabara, S., and Coauthors, 2016: S4: An O2R/R2O infrastructure for optimizing satellite data utilization in NOAA numerical modeling systems: A step toward bridging the gap between research and operations. *Bull. Amer. Meteor. Soc.*, **97**, 2359–2378, doi:10.1175/BAMS-D-14-00188.1.

Buehner, M., 2005: Ensemble-derived stationary and flow-dependent background-error covariances: Evaluation in a quasi-operational NWP setting. *Quart. J. Roy. Meteor. Soc.*, **131**, 1013–1043, doi:10.1256/qj.04.15.

Cardinali, C., 2009: Monitoring the observation impact on the short-range forecast. *Quart. J. Roy. Meteor. Soc.*, **135**, 239–250, doi:10.1002/qj.366.

Carrassi, A., A. Trevisan, and F. Uboldi, 2007: Adaptive observations and assimilation in the unstable subspace by breeding on the data-assimilation system. *Tellus*, **59A**, 101–113, doi:10.1111/j.1600-0870.2006.00210.x.

Clayton, A., A. Lorenc, and D. Barker, 2013: Operational implementation of a hybrid ensemble/4D-Var global data assimilation system at the Met Office. *Quart. J. Roy. Meteor. Soc.*, **139**, 1445–1461, doi:10.1002/qj.2054.

Daescu, D. N., 2009: On the deterministic observation impact guidance: A geometrical perspective. *Mon. Wea. Rev.*, **137**, 3567–3574, doi:10.1175/2009MWR2954.1.

Ehrendorfer, M., R. M. Errico, and K. D. Raeder, 1999: Singular-vector perturbation growth in a primitive equation model with moist physics. *J. Atmos. Sci.*, **56**, 1627–1648, doi:10.1175/1520-0469(1999)056<1627:SVPGIA>2.0.CO;2.

Gaspari, G., and S. E. Cohn, 1999: Construction of correlation functions in two and three dimensions. *Quart. J. Roy. Meteor. Soc.*, **125**, 723–757, doi:10.1002/qj.49712555417.

Gasperoni, N. A., and X. Wang, 2015: Adaptive localization for the ensemble-based observation impact estimate using regression confidence factors. *Mon. Wea. Rev.*, **143**, 1981–2000, doi:10.1175/MWR-D-14-00272.1.

Geer, A., 2016: Significance of changes in medium-range forecast scores. *Tellus*, **68A**, 30 229, doi:10.3402/tellusa.v68.30229.

Gelaro, R., and Y. Zhu, 2009: Examination of observation impacts derived from observing system experiments (OSEs) and adjoint models. *Tellus*, **61A**, 179–193, doi:10.1111/j.1600-0870.2008.00388.x.

——, R. H. Langland, S. Pellerin, and R. Todling, 2010: The THORPEX Observation Impact Intercomparison Experiment. *Mon. Wea. Rev.*, **138**, 4009–4025, doi:10.1175/2010MWR3393.1.

Holdaway, D., R. Errico, R. Gelaro, and J. G. Kim, 2014: Inclusion of linearized moist physics in NASA's Goddard Earth Observing System data assimilation tools. *Mon. Wea. Rev.*, **142**, 414–433, doi:10.1175/MWR-D-13-00193.1.

Hunt, B. R., E. J. Kostelich, and I. Szunyogh, 2007: Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D*, **230**, 112–126, doi:10.1016/j.physd.2006.11.008.

Ingleby, N. B., and A. C. Lorenc, 1993: Bayesian quality control using multivariate normal distribution. *Quart. J. Roy. Meteor. Soc.*, **119**, 1195–1225, doi:10.1002/qj.49711951316.

Isaksen, L., M. Fisher, E. Andersson, and J. Barkmeijer, 2005: The structure and realism of sensitivity perturbations and their interpretation as 'Key Analysis Errors.' *Quart. J. Roy. Meteor. Soc.*, **131**, 3053–3078, doi:10.1256/qj.04.99.

Ishibashi, T., 2010: Optimization of error covariance matrices and estimation of observation data impact in the JMA global 4D-Var system. CASJSC WGNE Research Activities in Atmospheric and Oceanic Modelling, World Climate Research Programme Rep. 40, 1–11.

Kalnay, E., 2003: *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, 341 pp.

——, Y. Ota, T. Miyoshi, and J. Liu, 2012: A simpler formulation of forecast sensitivity to observations: application to ensemble Kalman filters. *Tellus*, **64A**, 18 462, doi:10.3402/tellusa.v64i0.18462.

Keyser, D., 2011: Observational data processing at NCEP. NOAA/NCEP/NWS/Environmental Modeling Center. [Available online at http://www.emc.ncep.noaa.gov/mmb/data_processing/data_processing/.]

——, 2013: PREPBUFR processing at NCEP. NOAA/NCEP/NWS/Environmental Modeling Center. [Available online at http://www.emc.ncep.noaa.gov/mmb/data_processing/prepbufr.doc/document.htm.]

Kleist, D. T., 2012: An evaluation of hybrid variational-ensemble data assimilation for the NCEP GFS. Ph.D. dissertation, University of Maryland, College Park, 149 pp. [Available online at http://hdl.handle.net/1903/13135.]

——, and M. C. Morgan, 2005: Application of adjoint-derived forecast sensitivities to the 24–25 January 2000 U.S. east coast snowstorm. *Mon. Wea. Rev.*, **133**, 3148–3175, doi:10.1175/MWR3023.1.

——, and K. Ide, 2015a: An OSSE-based evaluation of hybrid variational–ensemble data assimilation for the NCEP GFS. Part I: System description and 3D-hybrid results. *Mon. Wea. Rev.*, **143**, 433–451, doi:10.1175/MWR-D-13-00351.1.

——, and ——, 2015b: An OSSE-based evaluation of hybrid variational–ensemble data assimilation for the NCEP GFS. Part II: 4DEnVar and hybrid variants. *Mon. Wea. Rev.*, **143**, 452–470, doi:10.1175/MWR-D-13-00350.1.

Kumar, K., J. C. Alpert, D. L. Carlis, and B. A. Ballish, 2009: Investigation of NCEP GFS model forecast skill "dropout" characteristics using the EBI index. *23rd Conf. on Weather Analysis and Forecasting/19th Conf. on Numerical Weather Prediction*, Omaha, NE, Amer. Meteor. Soc., 13A.1. [Available online at https://ams.confex.com/ams/23WAF19NWP/techprogram/paper_154282.htm.]

Langland, R. H., and N. L. Baker, 2004: Estimation of observation impact using the NRL atmospheric variational data assimilation adjoint system. *Tellus*, **56A**, 189–201, doi:10.3402/tellusa.v56i3.14413.

Li, H., J. Liu, and E. Kalnay, 2010: Correction of "Estimating observation impact without adjoint model in an ensemble Kalman filter." *Quart. J. Roy. Meteor. Soc.*, **136**, 1652–1654, doi:10.1002/qj.658.

Lien, G.-Y., 2014: Ensemble assimilation of global large-scale precipitation. Ph.D. dissertation, University of Maryland, College Park, 165 pp. [Available online at http://hdl.handle.net/1903/15274.]

Liu, J., and E. Kalnay, 2008: Estimating observation impact without adjoint model in an ensemble Kalman filter. *Quart. J. Roy. Meteor. Soc.*, **134**, 1327–1335, doi:10.1002/qj.280.

——, ——, T. Miyoshi, and C. Cardinali, 2009: Analysis sensitivity calculation in an ensemble Kalman filter. *Quart. J. Roy. Meteor. Soc.*, **135**, 1842–1851, doi:10.1002/qj.511.

Lorenc, A. C., 1981: A global three-dimensional multivariate statistical interpolation scheme. *Mon. Wea. Rev.*, **109**, 701–721, doi:10.1175/1520-0493(1981)109<0701:AGTDMS>2.0.CO;2.

——, and R. T. Marriott, 2014: Forecast sensitivity to observations in the Met Office global numerical weather prediction system. *Quart. J. Roy. Meteor. Soc.*, **140**, 209–223, doi:10.1002/qj.2122.

Onogi, K., 1998: A data quality control method using forecasted horizontal gradient and tendency in a NWP system: Dynamic QC. *J. Meteor. Soc. Japan*, **76**, 497–516, doi:10.2151/jmsj1965.76.4_497.

Ota, Y., J. C. Derber, T. Miyoshi, and E. Kalnay, 2013: Ensemble-based observation impact estimates using the NCEP GFS. *Tellus*, **65A**, 20 038, doi:10.3402/tellusa.v65i0.20038.

Pires, C., R. Vautard, and O. Talagrand, 1996: On extending the limits of variational assimilation in nonlinear chaotic systems. *Tellus*, **48A**, 96–121, doi:10.3402/tellusa.v48i1.11634.

Rodwell, M. J., and Coauthors, 2013: Characteristics of occasional poor medium-range weather forecasts for Europe. *Bull. Amer. Meteor. Soc.*, **94**, 1393–1405, doi:10.1175/BAMS-D-12-00099.1.

Simmons, A., 2011: From observations to service delivery: Challenges and opportunities. *WMO Bulletin*, Vol. 60, No. 2, WMO, Geneva, Switzerland, 96–107. [Available online at https://public.wmo.int/en/bulletin/observations-service-delivery-challenges-and-opportunities.]

Tavolato, C., and L. Isaksen, 2015: On the use of a Huber norm for observation quality control in the ECMWF 4D-Var. *Quart. J. Roy. Meteor. Soc.*, **141**, 1514–1526, doi:10.1002/qj.2440.

Trevisan, A., and F. Uboldi, 2004: Assimilation of standard and targeted observations within the unstable subspace of the observation–analysis–forecast cycle system. *J. Atmos. Sci.*, **61**, 103–113, doi:10.1175/1520-0469(2004)061<0103:AOSATO>2.0.CO;2.

——, M. D'Isidoro, and O. Talagrand, 2010: Four-dimensional variational assimilation in the unstable subspace and the optimal subspace dimension. *Quart. J. Roy. Meteor. Soc.*, **136**, 487–496, doi:10.1002/qj.571.

Uboldi, F., and A. Trevisan, 2006: Detecting unstable structures and controlling error growth by assimilation of standard and adaptive observations in a primitive equation ocean model. *Nonlinear Processes Geophys.*, **13**, 67–81, doi:10.5194/npg-13-67-2006.

Wang, X., D. Parrish, D. Kleist, and J. Whitaker, 2013: GSI 3DVar-based ensemble-variational hybrid data assimilation for NCEP Global Forecast System: Single-resolution experiments. *Mon. Wea. Rev.*, **141**, 4098–4117, doi:10.1175/MWR-D-12-00141.1.

Whitaker, J. S., and T. M. Hamill, 2002: Ensemble data assimilation without perturbed observations. *Mon. Wea. Rev.*, **130**, 1913–1924, doi:10.1175/1520-0493(2002)130<1913:EDAWPO>2.0.CO;2.