Proactive Quality Control: Observing System Simulation Experiments with the Lorenz '96 Model

TSE-CHUN CHEN AND EUGENIA KALNAY

Department of Atmospheric and Oceanic Science, University of Maryland, College Park, College Park, Maryland

(Manuscript received 17 April 2018, in final form 20 October 2018)

ABSTRACT

Proactive quality control (PQC) is a fully flow-dependent QC for observations based on the ensemble forecast sensitivity to observations technique (EFSO). It aims at reducing the forecast skill dropout events suffered in operational numerical weather prediction by rejecting observations identified as detrimental by EFSO. Past studies show that individual dropout cases from the Global Forecast System (GFS) were significantly improved by noncycling PQC. In this paper, we perform for the first time cycling PQC experiments in a controlled environment with the Lorenz model to provide a systematic testing of the new method and possibly shed light on the optimal configuration of operational implementation. We compare several configurations and PQC update methods. It is found that PQC improvement is insensitive to the suboptimal configurations in DA, including ensemble size, observing network size, model error, and the length of DA window, but the improvements increase with the flaws in observations. More importantly, we show that PQC improves the analysis and forecast even in the absence of flawed observations. The study reveals that reusing the exact same Kalman gain matrix for PQC update not only provides the best result but requires the lowest computational cost among all the tested methods.

1. Introduction

Modern data assimilation (DA) in operational numerical weather prediction (NWP) has evolved into complex systems that ingest high volumes of data of many classes, prohibiting the detailed evaluation of the contributions from small subsets of observations through the use of data denial experiments (observing system experiments). To address this problem, important studies have focused on a more efficient estimation of observational impact using forecast sensitivity to observations. Langland and Baker (2004) first developed the adjoint-based forecast sensitivity to observation (adjoint FSO). With the adjoint model, FSO propagates back (to observation time) future forecast error changes between two consecutive forecasts resulting from data assimilation and attributes those changes to each individual observation, allowing detailed observational impact estimation for the first time. Kalnay et al. (2012) developed an ensemble equivalent of the adjoint FSO (EFSO), replacing the adjoint model with the use of ensemble forecasts, and shed light on the estimation of observational impact without the need of an adjoint model. In addition, Buehner et al. (2018) formulated an ensemble–variational (EnVar) approach that computes the observational impact through minimization of a cost function as in the adjoint-based approach but uses the ensemble when the adjoint model is needed. To date, many operational centers have adopted at least one of the approaches of FSO for a better understanding of the complex DA systems (Zhu and Gelaro 2008; Cardinali 2009; Gelaro et al. 2010; Lorenc and Marriott 2014; Ota et al. 2013; Sommer and Weissmann 2014).

Several studies have investigated the application of (generic) FSO impacts. Lien et al. (2018) demonstrated with an example of precipitation assimilation that using long-term averaged noncycled EFSO impact as guidance can accelerate the development of data selection and quality control procedures for new observing systems. Kotsuki et al. (2017) found that among all the ordering methods tested on a serial ensemble square root filter (EnSRF), the analysis can be significantly improved by ordering the observations from detrimental to beneficial based on EFSO. Proactive quality control (PQC; Ota et al. 2013; Hotta et al. 2017a) based on EFSO was proposed as a strategy...
for reducing dropouts in forecast skill (Kumar et al. 2017) through identification and rejection of detrimental observations that may be harmful to the forecast. Ota et al. (2013) showed using the Global Forecast System (GFS) from the National Centers for Environmental Prediction (NCEP) that denying the detrimental observations identified by EFSo with a 24-h verification lead time reduced forecast errors in several forecast skill dropout cases. Hotta et al. (2017a) successfully showed with 20 forecast skill dropout cases that the forecast errors can be reduced by data denial based on EFSo with lead time of only 6 h. Furthermore, it was proposed that PQC would be more beneficial for sequential (cycling) data assimilation and that it is affordable in NCEP operation with some shortcuts (Hotta et al. 2017a; Chen 2018). A major potential benefit in cycling PQC is that the improved forecast may serve as a better background and subsequently lead to improvement in the following analyses and forecasts. However, cycling PQC has not yet been thoroughly tested. Idealized simulation experiments in a controlled environment can provide insights on how to optimally set up cycling PQC for realistic models.

In this study, we implement EFSo and PQC on the simple Lorenz (1996) model with 40 variables coupled with the ensemble transform Kalman filter (ETKF; Bishop et al. 2001) for DA. This ideally controlled environment allows the separation of factors contributing to the errors in DA, EFSo, and PQC that are entangled together in realistic systems. Section 2 reviews briefly the ETKF and EFSo formulation and introduces the PQC algorithm together with several proposed PQC update methods. The experimental setup will be described in section 3. Section 4 shows the results of the PQC performance using various configurations as well as its sensitivity to a suboptimal DA system. We summarize the findings of the study in section 5.

2. Methodology

a. ETKF

The ensemble Kalman filter (EnKF) is one of the prevailing methods for data assimilation, combining a model forecast with observations to construct a linear least squares error estimation of the true state or analysis. An ensemble of simulations initiated from K perturbed states, forms the flow-dependent error covariance of the forecast (also known as background error covariance

\[ \mathbf{B} = \frac{1}{(K-1)} \mathbf{X}^T \mathbf{X}, \]

where \( \mathbf{X} \) is the background perturbation matrix, whose columns are the forecast ensemble perturbations with respect to the mean \( \mathbf{x}^f \). The background error covariance \( \mathbf{B} \) accounts for model uncertainty and cross-variable correlations. It is then used together with the observational error covariance \( \mathbf{R} \) to combine the background state \( \mathbf{x}^f \) and the observations \( \mathbf{y} \) into the analysis state \( \mathbf{x}^a \). The analysis equation can be written as

\[ \mathbf{x}^a = \mathbf{x}^f + \mathbf{K}(\mathbf{y} - \mathbf{Hx}^f), \]

\[ \mathbf{K} = (\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{R}^{-1}, \]

where \( \mathbf{A} = 1/(K-1)\mathbf{X}^T \mathbf{x}^f \) represents the analysis error covariance, \( \mathbf{X}^T \) represents the analysis perturbations, and \( \mathbf{H} \) represents the observation operator.

b. EFSo formulation

EFSo is the ensemble version of the adjoint FSo that replaces the adjoint model with ensemble forecasts to propagate the forecast sensitivity back to the observation time and attribute them to each observation based on their contributions to analysis increment (AI).

We introduce the formulation of EFSo following Kalnay et al. (2012). The forecast errors for the same verification time \( t \) initiated from 0- and 6-h analyses are, respectively, denoted as

\[ e_{x/0}^f = x_{x/0}^f - x_t^f, \]

\[ e_{x/-6}^f = x_{x/-6}^f - x_t^f, \]

where \( x_{x/0}^f \) and \( x_{x/-6}^f \) represent the forecasts valid at time \( t \) initiated from time 0 and -6, respectively, and \( x_t^f \) is the verifying truth for the forecast at time \( t \). In practice, the best option available for \( x_t^f \) in real time is to use \( x_t^f \), the analysis at time \( t \).

The difference between the two forecast errors is introduced by the data assimilation at time 0, and the forecast error changes can be measured as...
\[ \Delta \varepsilon^2 = e_{t_0}^T \mathbf{C}_t e_{t_0} - e_{t_1}^T \mathbf{C}_t e_{t_1} \]
\[ = (e_{t_0} - e_{t_1})^T \mathbf{C}_t (e_{t_0} + e_{t_1}) \]
\[ = (x_{t_0} - x_{t_1})^T \mathbf{C}_t (e_{t_0} + e_{t_1}). \] (8)

where \( \mathbf{C} \) is the chosen error norm matrix. By using the Jacobian of the 6-h forecast operator \( \mathbf{M} \) and Eq. (1), we can approximate the forecast error change in Eq. (8) as
\[ \Delta \varepsilon^2 \approx \left[ \mathbf{M} (x_{t_0} - x_{t_1}) \right]^T \mathbf{C}_t (e_{t_0} + e_{t_1}) \], (9)
\[ = [\mathbf{M} \mathbf{K} \delta y_{t_0}]^T \mathbf{C}_t (e_{t_0} + e_{t_1}), \] (10)
\[ = \delta y_{t_0}^T \mathbf{K}^T \mathbf{M}^T \mathbf{C}_t (e_{t_0} + e_{t_1}). \] (11)

Substituting Eq. (2) and the approximation \( \mathbf{H} x_{t_0}^a \approx y_{t_0}^a \) into Eq. (11) gives
\[ \Delta \varepsilon^2 \approx \frac{1}{K-1} \delta y_{t_0}^T \mathbf{R}^{-1} y_{t_0}^a \mathbf{X}_{t_0}^a \mathbf{C}_t (e_{t_0} + e_{t_1}). \] (12)

where \( y_{t_0}^a \) represents the analysis perturbation in the observation space, and \( \mathbf{X}_{t_0}^a \) is the background perturbation initiated from time 0 and valid at time \( t \).

The impact of each observation can then be obtained by decomposing the sum of the inner product of the innovation vector \( \delta y_0 \) and the sensitivity vector \( \left[(1/(K-1)) \mathbf{R}^{-1} y_{t_0}^a \mathbf{X}_{t_0}^a \mathbf{C}_t (e_{t_0} + e_{t_1}) \right] \) into elements that correspond to each observation, so that \( \left[(1/(K-1)) \delta y_{t_0}^T \mathbf{R}^{-1} y_{t_0}^a \mathbf{X}_{t_0}^a \mathbf{C}_t (e_{t_0} + e_{t_1}) \right] \) represents the impact of the \( t \)th observation on the forecast error changes. This impact represents the estimated forecast error changes caused by the assimilation of the corresponding observation. Positive (negative) impact values indicate that the forecast error increases (reduces) by assimilating such observations, and hence, we call them detrimental (beneficial) observations. This terminology was introduced in Hotta et al. (2017a) to avoid the confusion between “positive EFSO impact value,” which indicates a detrimental impact, with “positive” (beneficial) impact.

c. Proactive quality control algorithm

The fundamental concept of PQC is to utilize the EFSO impact as observational QC in each DA cycle for identifying detrimental observations. The analysis is then modified to avoid the impact of those detrimental observations. It should be noted that EFSO cannot be computed until the next analysis becomes available for forecast error verification. The PQC algorithm is summarized in Fig. 1. It inserts additional steps (verifying analysis for EFSO, EFSO computation, PQC analysis, update, and the forecast from the updated analysis) into a standard DA cycle. A major focus of this study is to compare the performance of five possible PQC analysis updates defined as follows (summarized in Table 1):

- **PQC_H**: This method avoids the influence of detrimental observations by deleting their corresponding columns from the observational operator \( \mathbf{H} \), which is equivalent to ignoring the observation. This is the method adopted in Ota et al. (2013) and Hotta et al. (2017a) for PQC. Since it is easy to implement, this method is also commonly used for data denial. A major disadvantage is that it is necessary to repeat the analysis step to obtain the PQC update.

- **PQC_R**: This method removes the influence of detrimental observations by increasing the corresponding observational error covariance \( \mathbf{R} \). It can be viewed as a soft version of PQC_H with the flexibility of tuning the magnitude of the detrimental influence. It is also commonly used for data denial and easy to implement. It has the same drawback as does PQC_H, requiring the repetition of the analysis step.

- **PQC_K**: This method was proposed in Ota et al. (2013) and Hotta et al. (2017a) as an approximation to PQC_H by computing the PQC correction to analysis using the exact same gain matrix \( \mathbf{K} \) but setting the detrimental innovations to zero. However, this method is actually more consistent with EFSO since it also assumes the gain matrix to be the same for impact estimation. Mathematically, the PQC_K corrected analysis:
\[ x_0^{PQC,K} = x_0^a - \mathbf{K} \begin{pmatrix} 0 \\ \delta y_0^{deny} \end{pmatrix} \]
\[ = x_0^j + \mathbf{K} \begin{pmatrix} \delta y_0 - (0 \delta y_0^{deny}) \end{pmatrix} . \]

where \( \delta y_0 \) represents the original innovation and the superscript “deny” denotes the rejected elements of the innovation vector. PQC_K does not require repeating the process of computing the gain matrix \( \mathbf{K} \) (analysis step) so that it significantly reduces the computational burden. In section 4b, we will show that this method actually performs much better than PQC_H and PQC_R.

- **PQC_BmO**: Another approach, following the thought of not repeating the analysis step and the idea of serial EnKF, is to treat the original analysis as background and assimilate the innovation [observation minus background (OmB)] associated with detrimental observations again but with the opposite sign (BmO),
thus canceling the influence of those observations. This can be expressed as follows:

\[
x_{0}^{\text{PQC_BmO}} = x_{0}^{b} - K_{\text{PQC}}(\delta y_{0}^{\text{deny}})
\]

\[
= x_{0}^{b} + (K - [0 \ K_{\text{PQC}}])\delta y_{0}^{\text{deny}}
\]

\[
K_{\text{PQC}} = (A^{-1} + H_{\text{PQC}}^{T}R_{\text{PQC}}^{-1}H_{\text{PQC}})^{-1}H_{\text{PQC}}^{T}R_{\text{PQC}}^{-1}
\]

where \(H_{\text{PQC}}\) and \(R_{\text{PQC}}\) is the same as the original observation operator and error covariance but only with the blocks associated with rejected observations.

- **PQC_AmO:** This method is a variant of the PQC_BmO. The only difference between the two is the definition of the innovation. Innovation in PQC_BmO is defined as the observation minus the original background while it is observation minus original analysis in PQC_AmO:

\[
x_{0}^{\text{PQC_AmO}} = x_{0}^{b} - K_{\text{PQC}}(y_{0}^{\text{deny}} - H_{\text{PQC}}x_{0}^{b})
\]

\[
= x_{0}^{b} + (K - [0 \ K_{\text{PQC}}(I - H_{\text{PQC}})K])\delta y_{0}^{\text{deny}},
\]

where \(y_{0}^{\text{deny}}\) represents the rejected observations and \(K\) denotes the corresponding block of the original \(K\).

The PQC methods differ mostly in computational requirement and the changes in \(K\). The PQC_H and the PQC_R require the largest computational resources, while PQC_K poses the lowest computational burden. In terms of the change in gain matrix \(K\), it is modified the most in POC_H, POC_R, and moderately in POC_BmO and POC_AmO. POC_K does not change the gain matrix \(K\) at all. We will discuss more on the importance of the change in \(K\) in section 4 when comparing the performance of the methods.

### 3. Experimental setup

The main purpose of this study is to test PQC for cycling data assimilation in an idealized simple system, which allows to clearly separate the impact of each factor, and to perform sensitivity tests more efficiently compared to a realistic system such as the GFS. To achieve this goal while remaining relevant to the realistic application in the atmosphere, we choose the one-dimensional simplified atmospheric model from Lorenz (1996), which resembles some of the large-scale atmospheric behavior and error growth characteristics. It is a model of \(N\) variables \(x_{1}, \ldots, x_{N}\) governed by \(N\) equations:

\[
\frac{dx_{n}}{dt} = x_{n-1}(x_{n+1} - x_{n-2}) - x_{n} + F,
\]

where the nonlinear terms on the right-hand side simulate advection while conserving total energy \(\sum_{n=1}^{N} x_{n}^{2}/2\), whereas \(-x_{n}\) and \(F\) represents dissipation and external forcing that drives the chaotic dynamics. We follow the commonly used configuration as in Lorenz and Emanuel (1998). The constant forcing term \(F\) is set to 8, so that the error doubling rate corresponding to the leading

### Table 1. Summary of PQC update methods.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Mechanism</th>
<th>Change in (K)</th>
<th>Change in Spread</th>
<th>Repeat analysis</th>
<th>Computational cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>PQC_H</td>
<td>Recompute (K) without observation</td>
<td>Large</td>
<td>Increased</td>
<td>Yes</td>
<td>High</td>
</tr>
<tr>
<td>PQC_R</td>
<td>Recompute (K) with upweighted (R)</td>
<td>Large</td>
<td>Increased</td>
<td>Yes</td>
<td>High</td>
</tr>
<tr>
<td>POC_K</td>
<td>Reuse the original EFSO (K)</td>
<td>None</td>
<td>None</td>
<td>No</td>
<td>Low</td>
</tr>
<tr>
<td>POC_BmO</td>
<td>Assimilate background minus observation</td>
<td>Low</td>
<td>Reduced</td>
<td>(Serial update)</td>
<td>Medium</td>
</tr>
<tr>
<td>POC_AmO</td>
<td>Assimilate analysis minus observation</td>
<td>Low</td>
<td>Reduced</td>
<td>(Serial update)</td>
<td>Medium</td>
</tr>
</tbody>
</table>
Lyapunov exponent $\lambda_1$ is about 0.42 model time units or 2.1 days, assuming that 0.05 model time unit is equivalent to 6 h in physical space based on the error growth rate. This time scale for error doubling is approximately consistent with that of the large-scale atmosphere in the midlatitudes. The time integration uses the fourth-order Runge–Kutta scheme with a time step of $\Delta t = 0.05$ model time units. We will be using this $\Delta t = 1$ step as the basic time unit throughout the paper instead of the commonly adopted conversion to physical 6 h since this can be misleading especially when the dynamical time scale is not necessarily relevant to the error-doubling time scale. As commonly done, the model dimension is chosen to be $N = 40$.

Each experiment is initialized from a randomly chosen state and spun up for 500 time steps, allowing the ensemble members to converge to the model attractor. For the control and PQC experiments, an additional 500-step spunup for DA is performed. Each experimental period is 5000-step long after spinup. A “truth” run without DA and PQC is performed in order to generate the observations and verify the performance of the experiments. Following Lorenz and Emanuel (1998), the observations are generated at $\Delta t = 0.05$ intervals by adding to the truth random observational noise drawn from normal distribution $N(\mu, R)$. Unless otherwise stated, observation errors are generated with mean $\mu = 0$ and standard deviation $R = 0.01$.

For “flawed” (imperfect) observing system experiments, we modify the values of $\mu$ and $R$ to make them inconsistent with the prescribed observational error covariance matrix $R$. The sensitivity to the spatial coverage of the observing network is also tested. If the number of observation locations is not equal to 40, their position is randomly chosen from a uniform distribution in each cycle.

We chose to perform ETKF with a perfect model and ensemble size of 40 members because we are interested in assessing the EFSO and PQC performance without the need for localization and inflation, which are designed to deal with insufficient ensemble sizes and model error in EnKF (Liu and Kalnay 2008). The performance of each experiment is evaluated by computing the root-mean-square difference between the ensemble mean forecast/analysis and the truth from the nature run. Note that EFSO (and hence PQC) are norm-dependent methods. In this study, the error norm $C$ is chosen to be the identity $I$, so that each grid point is equally weighted. Other norms can be used for different purposes.

4. Proactive QC

In section 4a, we examine the performance of noncycling PQC in which the improved forecast is not used as background for the following analysis, as in Hotta et al. (2017a). In section 4b, the first experimental cycling of PQC is performed. We compare all the PQC methods introduced in section 2, each using different mechanisms to avoid the impact of the detrimental observations. In section 4c, the sensitivity of PQC performance to the choice of EFSO lead time and the amount of rejected observations are investigated. Finally, the robustness of PQC is tested in section 4d by introducing different sources of imperfections in the DA system, all relevant to operational applications.

a. Noncycling PQC

In this subsection, we explore the performance of PQC in noncycling fashion as in Hotta et al. (2017a). One of the key highlights of the study was the design of the data denial strategy. The EFSO impact was used as a guide on the denying priority, but not every detrimental observation was rejected. Note that only slightly more than 50% of the observations are beneficial on average, as reported by every operational center and research group, and it has been generally believed that rejecting nearly 50% of the observations would lead to forecast degradation. Additionally, every observation, by the nature of DA, provides an extra piece of information that should be useful in estimating the true state of the trajectory. Given the advantages of this simple low-dimensional system, we are able to test the sensitivity of the PQC performance to the number of rejected observations, with great granularity. In Fig. 2, we show how the noncycling PQC_H improves or degrades the forecasts by varying the number of rejected observations ordered from the most detrimental to the most beneficial. In this case, all the observations are perfectly consistent with the prescribed error covariance $R$, meaning none of them are flawed. We observe a strong forecast error reduction resulting from rejecting the four most detrimental observations. Then the error reduction saturates when rejecting additional observations. And not until we reject the last four (most beneficial) observations, do we observe that a rapid error growth takes place. The error reduction by PQC is well preserved in time, and even amplifies in magnitude as the forecasts advance (note the vertical axis is in log scale). It should be noted here that in a noncycling PQC, the impact of using different EFSO lead times is not significant (not shown). This is consistent with the intuition that different lead times simply perturb the rejecting order slightly in most cases and only make significant changes in rare situations, so that their effect can only be observable after cycles of accumulation. This result is supportive evidence that EFSO can identify the very detrimental observations from a pool of observations. In addition,
the result confirms that rejecting few very detrimental observations provides the major contribution to the PQC correction. We note that all the proposed PQC update methods were tested with the noncycling experiments, and their results are almost identical to those using the PQC_H method.

b. Cycling PQC: Methods

Figure 3 compares the performance of all proposed PQC methods, using six time steps as a function of rejecting percentages of observations. Note that the magnitude of observational impacts are different and the percentages of beneficial observations could vary from 30% to 70% from cycle to cycle. There could be fewer observations with large detrimental impact for some of the cycles and more for others. Hence, it is not desirable to reject the same amount of observations throughout each cycle. Here we construct a range of thresholds corresponding to the 0th–100th percentiles of EFSO impacts obtained from a control experiment of 5000 cycles (shown in Table 2). Then PQC rejects observations with an impact above the threshold, where the 10th percentile threshold rejects the top 10% of the most detrimental observations and the 90th percentile threshold keeps only the top 10% of the most beneficial observations. The lead time here is chosen to be six steps rather arbitrarily, and the sensitivity to the lead time will be examined later in section 4c. Since the Kalman gain $K$ of PQC_R approaches that of PQC_H asymptotically with increasing observational error, it is not surprising to see that PQC_H and PQC_R methods perform more or less the same in terms of both the analysis and the 30-step forecast error reduction. The errors are reduced the most when rejecting 10% of the observations for the two methods. This result is consistent with the one from the noncycling experiment where the 10th percentile threshold rejects around four observations on average. It is somewhat surprising that PQC_K, PQC_BmO, and PQC_AmO all outperform PQC_H and PQC_R, which are the two most commonly used data denial methods. For the analysis quality improvement, the obvious choice of the threshold shifts toward 20%. PQC_K does not show any degradation of analysis until rejecting more than 60% of the observations, whereas PQC_BmO and PQC_AmO stop showing improvement after 50% and even suffer from filter divergence beyond 60%. For the forecast quality improvement, the dependence of PQC_BmO and PQC_AmO on the thresholds are qualitatively similar to that in analysis performance. It is quite shocking to find that PQC_K has nearly no dependence on the thresholds between the 10th and 60th percentile, especially when compared to the 10% optimal choice for PQC_H and PQC_R.

Intuitively, the “flat bottom” feature of PQC_K (rather than the “check mark” shape of PQC_H and

![Figure 2. Forecast RMSE from single cycle PQC_H as a function of the number of rejected observations. The colored lines represent the forecast error evolution from PQC analysis corresponding to increasing the number of rejected observations from red to blue. Different colors simply distinguish individual experiments with different numbers of rejected observations. The black lines mark the forecast error at 1, 10, 20, and 30 steps.](image1)

![Figure 3. Performance of 6-step POC with all 5 methods in terms of (a) analysis RMSE and (b) 30-step forecast RMSE as a function of rejection percentage.](image2)
PQC_R) is more consistent with the estimated impact of the observations since the magnitude of the impacts between the 10th and 60th percentile is really small compared to that of those below 10th percentile. And hence the error reduction should be insensitive (flat bottom) to the rejection of those observations between the 10th and 60th percentile. This explains why the results are better for PQC_K than for PQC_H since PQC_K is more consistent with the nature of EFSO and the estimated impact. Note that EFSO is simply an ensemble-based linear mapping between forecast error changes and the observational innovations, which is in turn associated with AIs through the gain matrix K. It provides the estimated impacts of each observation in the presence of all other assimilated observations, and hence the impacts remain valid as long as K does not change much. However, PQC_H and PQC_R significantly change K when rejecting more observations, thereby the accuracy of the estimated EFSO impacts becomes lower, and the PQC based on those impacts does not work as desired. The total AI obtained at the end of the DA update consists of the AIs contributed from each individual observation, and it is the AIs that determine the forecast error changes rather than the observation innovation. Hence, PQC should not target the detrimental observations themselves but the corresponding original AIs. Simple data denial experiments by manipulating H and R (PQC_H and PQC_R) reject the detrimental observations, but not the exact detrimental AIs as does PQC_K, especially when rejecting an excessive number of observations. PQC_K, by contrast, uses the exact same K to reject the exact detrimental AIs identified by EFSO and ends up with even larger improvements. Interestingly, PQC_K was originally proposed as an “approximation” of PQC_H in order to avoid the large computational cost of recomputing the analysis in realistic system applications (Hotta et al. 2017a; Ota et al. 2013). In reality, as discussed above, PQC_K is actually much more accurate in the context of PQC based on EFSO. In addition, the observations with the largest impacts contribute AIs that project among the most unstable modes, while the less impacting observations are associated with the neutral and stable modes which have little or no error growth. Hence, after rejecting the few very detrimental AIs, it does not matter much whether those less impactful AIs are rejected since the difference is very unlikely to grow in the future, thereby showing the flat bottom feature in the center of Fig. 3.

From the covariance perspective, PQC_K has another advantage over PQC_H and PQC_R. By using the same K, PQC_K maintains the ensemble spread, whereas the other two methods constantly overestimate the spread and hence the uncertainty of the mean state. This overestimation results in underweighting the background, which is actually more accurate, in the following cycles, and could also explain why rejecting more observations leads to error growth. It is worth noting that the commonly observed difference in the impact estimated by EFSO and observing system experiments/data denial experiments corresponds to the difference in PQC_K and PQC_H. In data denial experiments and PQC_H, the Kalman gain K is modified by changing the assimilated observations, whereas the original K is used in EFSO and PQC_K.

Finally, PQC_BmO and PQC_AmO change K in a less radical fashion by “assimilation” of pseudo-observations with opposite OmB and OmA using the original analysis as background, and yield improvements similar to PQC_K with a small number of rejected observations. But they suffer from filter divergence easily with a large number of rejected observations since the ensemble becomes overly confident because of the “additional” assimilation of opposite innovations.

By comparing the performance of PQC_K with other methods, we conclude that PQC_K rejects the detrimental part of AIs in unstable modes via observational space. In this regard, PQC can be viewed as being related to “key analysis errors” method that perturbs the initial condition based on the adjoint linear sensitivity to future forecast error (Klinker et al. 1998; Isaksen et al. 2005). However, there are fundamental differences between the two. EFSO (and hence PQC) linearly maps forecast error changes, which is the result of nonlinear model propagation of AIs, back to individual observations and the associated AIs, whereas the “key analysis error” identifies perturbations in the initial condition that could potentially reduce the future forecast error. Additionally, since the norm of PQC corrections is bounded by the difference between the background and

<table>
<thead>
<tr>
<th>Lead time (steps)</th>
<th>00th</th>
<th>20th</th>
<th>40th</th>
<th>60th</th>
<th>80th</th>
<th>100th</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>$4 \times 10^{-4}$</td>
<td>$3 \times 10^{-5}$</td>
<td>$3 \times 10^{-5}$</td>
<td>$-7 \times 10^{-5}$</td>
<td>$-5 \times 10^{-4}$</td>
<td>$-6 \times 10^{-1}$</td>
</tr>
<tr>
<td>11</td>
<td>$3 \times 10^{0}$</td>
<td>$1 \times 10^{-3}$</td>
<td>$9 \times 10^{-5}$</td>
<td>$-2 \times 10^{-4}$</td>
<td>$-1 \times 10^{-3}$</td>
<td>$-5 \times 10^{0}$</td>
</tr>
<tr>
<td>16</td>
<td>$3 \times 10^{1}$</td>
<td>$3 \times 10^{-3}$</td>
<td>$2 \times 10^{-4}$</td>
<td>$-4 \times 10^{-4}$</td>
<td>$-4 \times 10^{-3}$</td>
<td>$-3 \times 10^{1}$</td>
</tr>
<tr>
<td>21</td>
<td>$5 \times 10^{2}$</td>
<td>$7 \times 10^{-3}$</td>
<td>$6 \times 10^{-4}$</td>
<td>$-9 \times 10^{-4}$</td>
<td>$-9 \times 10^{-3}$</td>
<td>$-6 \times 10^{1}$</td>
</tr>
</tbody>
</table>

Table 2. Rejecting thresholds based on EFSO impact (J kg$^{-1}$) as a function of EFSO lead time and rejecting percentile.
the analysis, PQC is free from the additional tuning of the size of the correction required in key analysis errors. To advance the understanding of PQC, we explore the relation between the spatial pattern of the AIs and the corresponding EFSO impact. An example from a typical cycle using 21-step PQC_K in a Lorenz (1996) with reduced dimension (with only 20 model grids for figure readability) is shown in Fig. 4. It is clear that the largest increments in (Fig. 4c) are associated with the large magnitudes in the error covariance (Fig. 4a). We further show that the most detrimental and beneficial directions are revealed when sorted with the EFSO impact (Fig. 4b). In Fig. 4d, the most detrimental increments (15th–19th) share a similar pattern in model grids, and the pattern is the same for the most beneficial increment (0th) but with an opposite sign. The example shows supportive evidence that PQC rejects observations associated with AIs in similar detrimental directions and controls the error growth in undesired directions.

c. Cycling PQC: Sensitivity to EFSO lead times

In this section, we explore the sensitivity of cycling PQC_H, PQC_K, and PQC_AmO to the rejecting threshold and, more importantly, to the EFSO lead time. We refer to PQC based on t-step EFSO as t-step PQC hereafter for simplicity.

Beginning with PQC_H shown in Fig. 5, it seems to be quite difficult to conclude the relation between the PQC_H performance and the length of lead time simply from the analysis error reduction. It is also counterintuitive to find that 21-step PQC_H performs worse than that with only 6 steps. However, the dependence of the performance on forecast error reduction can be easily summarized as follows. The forecast quality increases with the lead times up to 16 steps and then remains the

![Fig. 4. (a) Analysis error covariance matrix from a typical cycle. (b) Sorted EFSO impact for all 40 assimilated observations. (c) Analysis increments on each model grid from each observation. (d) As in (c), but the order of the observations is sorted with EFSO impact from the most beneficial to the most detrimental to align with (b).](image)
same with 21 steps. The result suggests the optimal choice of EFSO lead time settles between 16 and 21 steps, best describing the impact corresponding to the underlying dynamical evolution. Short lead-time PQC seemingly reduces the analysis or even short-term forecast error but may be irrelevant to long-term error growth. This speculation can be confirmed by comparing the performances of 6-step and 21-step PQC_H. It is clear that a considerable portion of error reduction by 6-step PQC_H lies within the stable subspace and decays over time. Additionally, the corrections of 21-step PQC for rejecting 10% of the observations does not reduce much of the total analysis error compared to other lead times, but it turns out that those are the most relevant to error growth in the long term, and rejecting them leads to huge forecast improvement, after 30 steps (Fig. 5b).

Now we show in Fig. 6 the sensitivity of the PQC_K performance to the lead time and the rejecting threshold. It is qualitatively consistent with the result of PQC_H, where the maximum forecast rather than analysis error reduction increases with lead time and saturates at 16–21 steps. There is a general feature in forecast error reduction shared among all lead times. The error drops dramatically when rejecting with 10th percentile followed by the flat bottom feature when increasing the rejecting percentile. Then the filter diverges when rejecting beyond a critical percentile. The only differences are the percentile where PQC_K leads to filter divergence and the magnitudes of error reduction. The critical threshold gradually approaches from the 80th percentile for 6 steps to the 50th percentile for 21 steps.

This dependence of PQC_K on the rejecting percentile is explained in Fig. 7, which is a schematic of EFSO computation in an one-dimensional model space with a group of unbiased observations centering around the truth and forecasts initiated from \(T = -6\) and \(T = 0\) lying somewhere with a distance to the truth. We can think of the one-dimensional model space in the figure being aligned with the fastest error growing subspace (first Lyapunov vector) as identified in Fig. 4d. And the error growth depends on the balance between the detrimental AIs and the beneficial AIs. The outermost observations and the corresponding AIs in detrimental

![Fig. 5. Comparison of cycling PQC_H performance using various EFSO lead time as a function of rejection percentage in terms of (a) analysis and (b) 30-step forecast analysis RMSE.](image1)

![Fig. 6. As in Fig. 5, but for PQC_K.](image2)
direction are the main drivers that deviate the forecast from the truth trajectory, whereas the beneficial observations draw the forecast toward the truth. Ideally, we should just reject the fast-growing detrimental AIs because we could end up with only the unstable growing “beneficial” AIs when more are rejected. Rejecting the top 10% detrimental observations that are associated with the fast-growing detrimental AIs provides most of the error reduction. In contrast, the rest of the detrimental observations are associated with nongrowing AIs, and rejection of them are less influential and result in the flat bottom feature. Once rejecting more than 50%, the average beneficial percentage, the unstable growing beneficial AIs take over and lead to error growth in the beneficial direction. Note that EFSO with a shorter lead time cannot differentiate between observations with small to medium impact. And the direct consequences of using a shorter lead time is a smaller error reduction caused by rejecting some of the beneficial AIs while keeping some of the detrimental ones unintentionally. This also means that when rejecting more than 50% of the observations, the remaining AIs are not entirely pointing to the beneficial direction and delay the occurrence of the divergence.

With Fig. 8, we show the PQC_AmO performance sensitivity to the choice of lead time and rejecting percentile. It is clear that it performs comparably well or even better than PQC_K when rejecting a small number of observations. This result is due to the extra contraction on the ensemble spread (or reduction in covariance) by “assimilating additional observations” reflecting that PQC updated mean state is more accurate than the original mean. However, it is the overcontraction in the spread that leads to filter divergence when rejecting additional observations.

The reason why the error reduction saturates at longer lead times is the linear assumption of EFSO. The increasing nonlinearity in the errors with longer lead times degrades the linear mapping between observation time and forecast verification time. This issue can be revealed in the correlation between the actual forecast error changes caused by DA and the ensemble-explained counterpart in EFSO computation. Table 3 shows this correlation decreases as the verifying lead time increases, indicating that the linear mapping cannot keep track of the actual impact well in long lead times.

d. PQC with suboptimal DA systems

To remain relevant to applications in an operational environment, we would like to explore PQC in suboptimal conditions including imperfect model, flawed
observing system, DA window, and various sizes of ensemble and observing network. For the rest of the paper, we will be using 6-step PQC, which also improves the quality of the forecast (though not as much as 21-step PQC), but is less computationally expensive.

In high-dimensional complex chaotic systems, we do not have the luxury of using a sufficiently large ensemble size because of the limited computational resources. It is important to examine the performance of PQC with different numbers of ensemble members. We tested a wide range of ensemble sizes (from 5 to 640), as shown in Fig. 9. The experiments with ensemble size less than 40 suffered, as expected, from filter divergence, since no localization was applied. Surprisingly, PQC is able to reduce the analysis error significantly even though the filter still diverges. With about 40 ensemble members, ETKF works well and PQC improves the quality of analysis as expected, whereas the analysis error shows a slight increase as the ensemble size doubles monotonically beyond 40. This is a characteristic unique to the family of ensemble square root type of filter and has been documented in the literature (e.g., Lawson and Hansen 2005; Ng et al. 2011). The prevailing explanation is the high probability of having ensemble outliers leading to ensemble clustering, which can be ameliorated by applying additional random rotation to the transform matrix, but it is not of our interest in this study. The PQC analysis error also increases with ensemble size, but is still smaller than the control error except for PQC_H with a size larger than 320.

In a “realistic” application, the state is only partially observed, and hence it is important to show whether PQC also benefits the system even when the size of the observing network does not match the size of the model. For the extreme case where only 5 observations are available (12.5% observed), PQC_H seems to degrade both the analyses and the forecasts while PQC_K and PQC_AmO still improve the forecasts. This result again shows that PQC operates on the AIs rather than the observation itself since PQC_H degrades the system by changing the gain matrix K radically. For the observing network with a size larger than 5, the improvements with PQC_K and PQC_AmO are rather similar, showing that PQC performs well with a wide range of observing network sizes. In Figs. 10c and 10d, we show the PQC performance sensitivity to the length of the DA window, which changes the nonlinearity of the model forecast and the validity of linear error growth assumptions. Both the analysis and the 30-step forecast RMSE are reduced by PQC although showing the same increase as the control with an increasing DA window.

So far, we have been using observations with errors consistent with R, meaning the observing system is flawless. However, in the real world, the observational error covariances are never truly known. It is worth noting here that an ensemble forecast sensitivity method was proposed recently that provides a way for fine-tuning R (EFSR; Hotta et al. 2017b). In addition to the inaccuracy in error covariances, observational bias may pose an even greater danger of degrading the filter performance. To examine PQC under the influences of flawed observing systems, additional random errors and biases not reflected in R are added to observations at the 10th and 30th grid points respectively in two set of experiments similar to those in Liu and Kalnay (2008). The average EFSO impact for each grid is shown in Fig. 11 for a range of flaw magnitudes. The flawed observations are successfully identified, suggesting that EFSO can also be used as data selection or QC method as described in Lien et al. (2018). Both the detrimental impact of the flawed grid and the beneficial impact of the neighboring grids increase with the magnitude of the flaw. Figures 12a–d summarize the responses of PQC to both types of the observational flaws, and it is clear that analysis and forecast error reduction by PQC increases with the magnitude of the flaws. For biases larger than 0.5 the filter suffer from divergence in the control but PQC stabilizes the filter for this bias, which is similar to what we observe in some other borderline cases (not shown).

Besides flawed observations, model error is another source of error that deserves special attention. We examine the response of the control and PQC by setting

<table>
<thead>
<tr>
<th>Lead time</th>
<th>6</th>
<th>11</th>
<th>16</th>
<th>21</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation</td>
<td>0.86</td>
<td>0.85</td>
<td>0.83</td>
<td>0.79</td>
</tr>
</tbody>
</table>

Table 3. Correlation coefficients between the actual (due to DA) and the ensemble explained forecast error changes (total EFSO impact) as a function of verifying lead time (steps).

Fig. 9. PQC performance in terms of analysis RMSE as a function of ensemble size.
the forcing term $F$ to be slightly different from the nature run, which is both the verifying truth and where the observations were drawn from. Figures 12e and 12f visualize both the control and PQC error in analysis and forecast increases with the model error that will eventually lead to filter divergence. It is shown that PQC improvements are almost invariant with respect to the model error.

In this section, we have shown that PQC improves the quality of analysis and the forecast even in suboptimal DA system. The improvement is even larger when the imperfections are originated from flawed observations. In the biased observation case, it is shown that PQC provides extra stability and avoids filter divergence. Additionally, we found that PQC still improves the analysis and forecast even in the case of suboptimal

![Fig. 10. Performance of PQC_H, PQC_K, and PQC_AmO in terms of (a),(c) analysis RMSE and (b),(d) 30-step forecast RMSE as a function of (top) the size of observing network and (bottom) the DA window.](image)

![Fig. 11. Mean EFSO impact for each observation throughout 5000 cycles of experiment that assimilates flawed observations: (a) observation at 10th grid point with various random error larger than specified in observational error covariance matrix $R$ and (b) observation at 30th grid point with various systematic bias.](image)
configurations in DA as long as the quality of verifying analysis is still more accurate than that of the forecast.

5. Summary and discussion

In this study, we explore the characteristics of different PQC methods and the sensitivity of PQC to various factors using a simple Lorenz (1996) model with ETKF as the data assimilation scheme.

We examine the performance of PQC_H, PQC_R, PQC_K, PQC_BmO, and PQC_AmO using various configurations of different EFSO lead times and rejecting percentiles. We show that PQC_H and PQC_R are suboptimal and computationally more expensive than other methods for repeating the analysis process again. While the PQC_AmO and PQC_BmO can easily lead to filter divergence because of overconfidence, we found that PQC_K has the best performance because it
rejects the analysis increments (AIs) contributed by detrimental observations without changing the original EFSO gain matrix. We find that in the Lorenz (1996) system, even in the absence of flawed observations, the PQC rejection of the 10% most detrimental observations significantly improves the forecasts. Between 10% and 50%, however, the rejection of more observations has little effect, and the forecasts deteriorate significantly beyond 50% rejection. The improvement of PQC increases with the forecast length, but it saturates around 16–20 time steps as a balance between capturing long-term error growth and the linear assumption in EFSO.

We believe the source of the PQC correction comes mainly from future information (the verifying analysis for EFSO). This smoother aspect of PQC may explain the large error reduction even in a perfect system, where no defects in observations, DA, and model exist. We also examine PQC performance for suboptimal DA setup with varying ensemble and observing network size, DA window, biased observation, random error inconsistent with R, and model error. The results show that all the suboptimality leads to degradation of the control analysis and forecast quality, but PQC still improves the quality even in the extreme cases of filter divergence. The improvement grows with the magnitude of the flaw in the observations, meaning that PQC corrects for the defects in the observations. Since this feature is not observed in a typical smoother, which does not change the quality of final analysis, further investigation on the smoother aspect of PQC is needed.


In summary, we have demonstrated the beneficial impact of applying PQC in Lorenz (1996) system. PQC improves the system in the presence of flawed observations, showing its potential of being a care-free and automated QC scheme on the fly with the DA system. More importantly, even in a flawless system, PQC still improves the quality of analysis and forecast by eliminating “detrimental” growing components of the analysis increment.

Acknowledgments. We thank the reviewers for valuable comments and suggestions. This work is based on T. Chen’s Ph. D. dissertation supported by a NESDIS/JPSS Proving Ground (PG) and a Risk Reduction (RR) CICS Grant.

REFERENCES

Kalnay, E., Y. Ota, T. Miyoshi, and J. Liu, 2012: A simpler formulation of forecast sensitivity to observations: Application to


